DEVELOPMENT OF VIRTUAL SCREENING AND *IN SILICO* BIOMARKER IDENTIFICATION MODEL FOR PHARMACEUTICAL AGENTS

ZHANG JINGXIAN

NATIONAL UNIVERSITY OF SINGAPORE

2012

Development of Virtual Screening and *In Silico* Biomarker Identification Model for Pharmaceutical Agents

ZHANG JINGXIAN

(*B.Sc. & M.Sc., Xiamen University*)

A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
DEPARTMENT OF PHARMACY
NATIONAL UNIVERSITY OF SINGAPORE

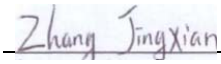
2012

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety.

I have duly acknowledged all the sources of information which have been used in the thesis

This thesis has also not been submitted for any degree in any university previously.



Zhang Jingxian

Acknowledgements

First and foremost, I would like to express my sincere and deep gratitude to my supervisor, Professor Chen Yu Zong, who gives me with the excellent guidance and invaluable advices and suggestions throughout my PhD study in National University of Singapore. Prof. Chen gives me a lot help and encouragement in my research as well as job-hunting in the final year. His inspiration, enthusiasm and commitment to science research greatly encourage me to become research scientist. I would like to appreciate him and give me best wishes to him and his loving family.

I am grateful to our BIDD group members for their insight suggestions and collaborations in my research work: Dr. Liu Xianghui, Dr. Ma Xiaohua, Dr. Jia Jia, Dr. Zhu Feng, Dr. Liu Xin, Dr. Shi Zhe, Mr. Han Bucong, Ms Wei Xiaona, Mr. Guo Yangfang, Mr. Tao Lin, Mr. Zhang Chen, Ms Qin Chu and other members. I honestly thank for their support for my research. It is a great honor to become a member of BIDD, which likes a big family. The great passion and successfulness of our BIDD group inspire me the most. I would also like to thank Prof. Yap Chun Wei, Prof. Guo Meiling for devoting their time as my QE examiners. I would like to thank Prof. Ji Zhiliang, my Master supervisor, for his great encouragement and help in my study in Xiamen and continue to support me in my PhD study and job hunting. I would like to thank Dr. Liu Xianghui for his great effort in teaching me in my research and warm invitations to his home. I would like to give my best wishes to him and his happy family. I would like to thank Dr. Wei Xiaona and Dr. Han Bucong for continuing encouragement and help in my research; I also like to give my best wishes to their future. I would also like to thank Mr. Wang Li, Mr. Li Fang, Mr. Wang Zhe and Mr. Patel Dhaval Kumar for their help in my study in pharmacy, I would like to wish them great future after graduation.

Lastly, I would like to thank my parents and my wife Gao Shizhen for their great cares on me all the time.

Zhang Jingxian, 2012

Table of Contents

Summary

Virtual screening (VS) especially machine learning based VS is increasingly used in search for novel lead compounds. It is a capable approach for facilitating hit lead compounds discovery. Various software tools have been developed for VS. However, conventional VS tools encounter issues such as insufficient coverage of compound diversity, high false positive rate and low speed in screening large compound libraries. Target selective drugs are developed for enhanced and reduced side effects. *In-silico* methods such as machine learning methods been explored for searching target selective ligands such as dopamine receptor ligands, but encountered difficulties associated with high subtype similarity and ligand structural diversity. In this thesis, we introduced a new two-step support vector machines target-binding and selectivity screening method for searching dopamine receptor subtype-selective ligands and demonstrated the usefulness of the new method in searching subtype selective ligands from large compound libraries. It has high subtype selective ligand identification rates as well as multi-subtype ligand identification rates. In addition, our method produced low false-hit rates in screening large compound libraries. Inhibitor of nuclear factor kappa-B (NF-κB) kinase subunit beta (IKKβ) has been a prime target for the development of NF-kB signaling inhibitors. In order to reduce the cost and time in developing novel IKKβ inhibitors, the machine learning method is used to build a prediction and screening model of IKKβ inhibitors. Our results show that support vector machine (SVM) based machine learning model has substantial capability in identifying IKKβ inhibitors at comparable yield and in many cases substantially lower false-hit rate than those of typical VS tools reported in the literatures and evaluated in this work. Moreover, it is capable of screening large compound

libraries at low false-hit rates.

Some drugs such as anticancer EGFR tyrosine kinase inhibitors elicit markedly different clinical response rates due to differences in drug bypass signaling as well as genetic variations of drug target and downstream drug-resistant genes. In this thesis, we systematically analyzed expression profiles together with the mutational, amplification and expression profiles of EGFR and drug-resistance related genes and investigated their usefulness as new sets of biomarkers for response of EGFR tyrosine kinase inhibitors. Our result shows that consideration of bypass signaling from pathway regulation perspectives appears to be highly useful for deriving knowledge-based drug response biomarkers to effectively predict drug responses well as for understanding the mechanism of pathway regulation and drug

List of Tables

List of Figures

List of Acronyms

VS	Virtual Screening
SBVS	Structure-based Virtual Screening
LBVS	Ligand-based Virtual Screening
ML	Machine Learning
P	Positive
N	Negative
kNN	k-nearest neighbors
MCC	Matthews correlation coefficient
PNN	Probabilistic neural network
SVM	Support vector machine
TP	True positive
TN	True negative
FP	False positive
FN	False negative
QSAR	Quantitative structure activity relationship
SAR	Structure-activity relationship
MCC	Matthews correlation coefficient
MDDR	MDL Drug Data Report
DR	Dopamine Receptor
RFE	Recursive Feature Elimination
Q	Overall Accuracy
IKKβ	Inhibitor of nuclear factor kappa-B kinase subunit beta
NFκB	Nuclear factor kappa-B kinase
EGFR	Epidermal growth factor receptor
TKI	Tyrosine kinase inhibitor
SVM-RFE	Support vector machine based recursive feature elimination
ADMET	Absorption, distribution, metabolism, excretion, toxicity

ANN	Artificial neural network
DI	Diversity index
CV	Cross validation

Chapter 1 Introduction

The process of new drugs discovery is normally a costly and time-consuming. The average time required for a successful drug development from initial design effort to market approval is about 13 years. Cheminformatics and bioinformatics tools are increasingly explored in facilitating pharmaceutical research and drug development. The thesis contains development of in silico virtual screening for potential pharmaceutical agents as well as discovery of biomarker for drug response. The introduction chapter includes: (1) Cheminiformatics in drug discovery (Section 1.1); (2) Cheminformatics and bioinformatics resources (Section 1.2); (3) Virutal screening of pharmaceutical agents (Section 1.3); (4) Bioinformatics tools in biomarker identification (Section 1.4); (5) Objectives and outlines (Section 1.5)

1.1 Cheminformatics in drug discovery

Traditionally, drug discovery process from idea to market consists of several steps: target discovery, lead compound screening, lead optimization, ADMET distribution, metabolism, excretion and toxicity) study, preclinical trial evaluation, clinical trials, and registration. It is a time-consuming, expensive, difficult, and inefficient process with low rate of new therapeutic discovery. The drug process takes approximately 10-17 years, $800 million (as per conservative estimates), the overall probability of success rate less than 10% [1] (**Figure 1-1**). The huge R&D investment in implementing new technologies for drug discovery does not

guarantee the increase of successful new chemical entities (NCEs). **Figure 1-2** shows the number of new chemical entities (NCEs) in relation to research and development (R&D) spending since 1992.

Figure 1-1 Drug discovery and development process (adopted from Ashburn et al. [1])

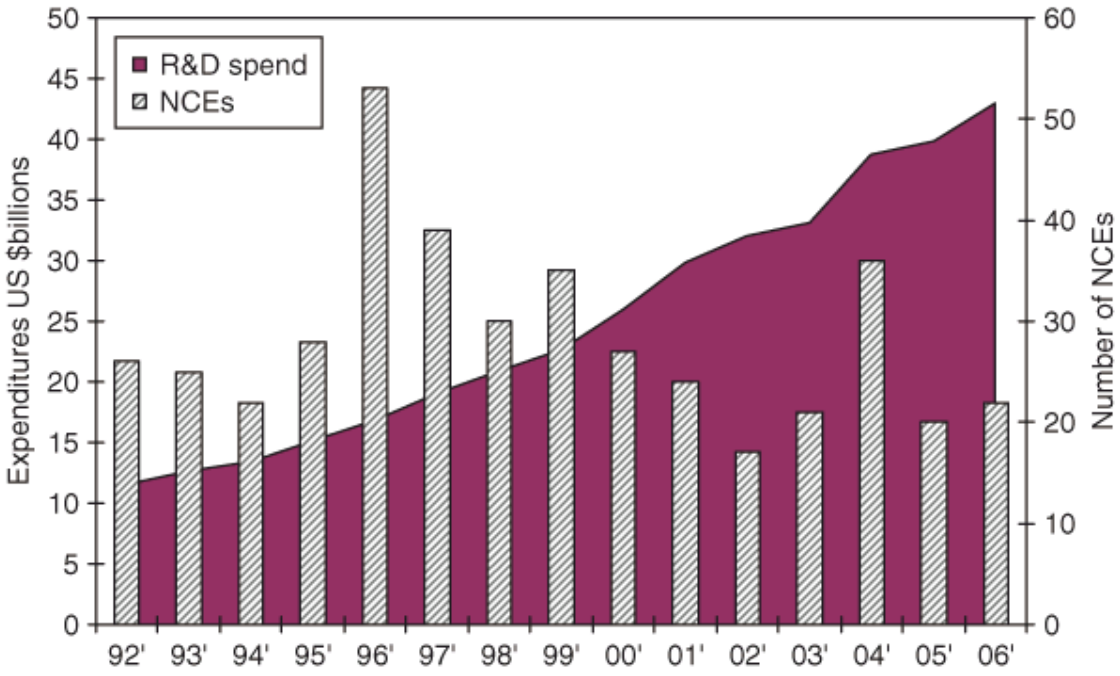


Figure 1-2 Number of new chemical entities (NCEs) in relation to research and development (R&D) spending (1992–2006). Source: Pharmaceutical Research and Manufacturers of America and the US Food and Drug Administration[2].

In order to increase the efficiency and reduce the cost and time of drug discovery, new technologies need to be employed in different stages of drug development

process. In particularly, earlier stages of drug discovery process, such as drug lead identification and optimization, toxicity of compounds estimation, are now greatly relying on new methodologies to reduce overall cost.

In 1990s, advances in the areas like molecular biology, cellular biology and genomics greatly help in understanding the molecular and genetic components in disease development and critical point in seeking therapeutic intervention. Technologies include DNA sequencing, microarray, HTS, combinatory chemistry, and high throughput sequencing have been developed. The progress is helpful in identifying many new molecular targets (from approximately 500 to more than 10,000 targets) [3]. In drug discovery, earlier stages, such as drug lead identification and optimization, toxicity of compounds estimation, are now greatly relying on new methodologies to reduce overall cost. High throughput screening (HTS) approaches for discovering potential therapeutic compounds on validated targets have been developed[4]. In the HTS process, compounds of diverse structure from chemical library are then screened against these validated targets[5]. Inspired by the terms genome and genomics after the finish of Human Genome Project, technologies such as motabolite profiles analysis and mRNA transcripts study that generate a lot of biological and chemistry data have been coined with the suffix -*ome* and –*omics*. **Table 1-1** lists a list of omics approaches and the fields they could be applied. The integration and annotation of the biological and chemical information to generate new knowledge become the major tasks of bioinformatics and cheminformatics.

Table 1-1 List of omics approaches and the fields they could be applied.

-ome	Fields of study (-omics)	Collection
Allergenome	Allergenomics	Proteomics of allergens
Bibliome	Bibliomics	Scientific bibliographic data
Connectome	Connectomics	Structural and functional brain connectivity at different spatiotemporal scales
Cytome	Cytomics	Cellular systems of an organism
Epigenome	Epigenomics	Epigenetic modifications
Exposome (2005)	Exposomics	An individual's environmental exposures, including in the prenatal environment
Exposome (2009)		Composite occupational exposures andoccupational health problems
Exome	Exomics	Exons in a genome
Genome	Genomics	Genes
Glycome	Glycomics	Glycans
Interferome	Interferomics	Interferons
Interactome	Interactomics	All interactions
Ionome	Ionomics	Inorganic biomolecules
Kinome	Kinomics	Kinases
Lipidome	Lipidomics	Lipids
Mechanome	Mechanomics	The mechanical systems within an organism
Metabolome	Metabolomics	Metabolites
Metagenome	Metagenomics	Genetic material found in an environmental sample
Metallome	Metallomics	Metals and metalloids
ORFeome	ORFeomics	Open reading frames (ORFs)
Organome	Organomics	Organ interactions
Pharmacogenetics	Pharmacogenetics	SNPs and their effect on pharmacokineticsand pharmacodynamics
Pharmacogenome	Pharmacogenomics	The effect of changes on the genome on pharmacology
Phenome	Phenomics	Phenotypes
Physiome	Physiomics	Physiology of an organism
Proteome	Proteomics	Proteins
Regulome	Regulomics	Transcription factors and other molecules involved in the regulation of gene expression
Secretome	Secretomics	Secreted proteins
Speechome	Speecheomics	Influences on language acquisition
Transcriptome	Transcriptomics	mRNA transcripts

According to the definition on Wikipedia, **Cheminformatics** is the use of computer and informational techniques, applied to a range of problems in the field of chemistry. Similarly, **bioinformatics** is the application of information

technology and computer science to the field of molecular biology. The main tasks that informatics handle are: to convert data to information and information to knowledge. According to market research firm BCC, the worldwide value of bioinformatics is increasing from $1.02 billion in 2002 to $3.0 billion in 2010, at an average annual growth rate (AAGR) of 15.8% (**Figure 1-3**). The use of bioinformatics in drug discovery is probably to cut the annual cost by 33%, and the time by 30% for developing a new drug. Bioinformatics and cheminformatics tools are getting developed which are capable to assemble all the required information regarding potential drug targets such as nucleotide and protein sequencing, homologue mapping[6, 7], function prediction[8, 9], pathway information[10], structural information[11] and disease associations[12], chemistry information.

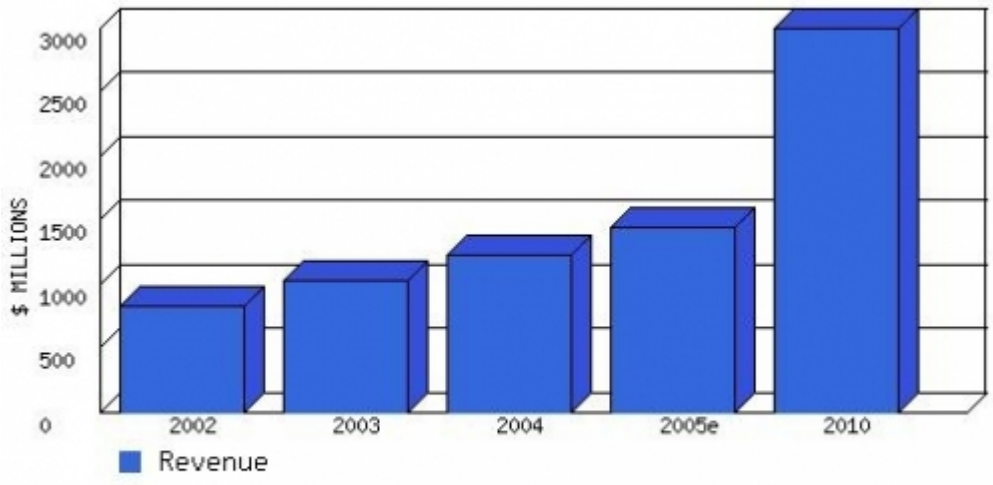


Figure 1-3 Worldwide value of bioinformatics Source: BCC Research[13]

1.2 Cheminformatics and bioinformatics resources

Currently there are many public bioinformatics databases (**Table 1-2**) and cheminformatics databases (**Appendix A Table 1**) that provide broad categories of medicinal chemicals, biomolecules or literature[14]. Bioinformatics databases mainly contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics. Information deposited in biological databases includes gene function, structure, clinical effects of mutations as well as similarities of biological sequences and structures. Cheminformatics database includes chemical and crystal structures, spectra, reactions and syntheses, and thermophysical data. For example, there are several known target and drug database including Drug Adverse Reaction Targets (DART), Therapeutic Target Database (TTD), Potential Drug Target Database (PDTD), PubChem, ChemblDB, BindingDB, DrugBank and etc.

Table 1-2 Popular bioinformatics database.

Database	Description
National Center for Biotechnology Information (NCBI) GenBank, EBI-EMBL, DNA Databank of Japan (DDBJ)	Databases with primary genomic data (complete genomes, plasmids, and protein sequences)
Swiss-Prot and TrEMBL and Protein Information Resource (PIR)	Databases with annotated protein sequences
COG/KOG (Clusters of Orthologous groups of proteins) and Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologies	Databases with results of cross-genome comparisons
Pfam and SUPFAM, and TIGRFAMs	Databases containing information on protein families and protein classification
TIGR Comprehensive Microbial Resource (CMR) and Microbial Genome Database for Comparative Analysis (MBGD)	Web services for cross-genome analysis
DIP, BIND, InterDom, and FusionDB	Databases on protein–protein interactions
KEGG and PathDB	Databases on metabolic and regulatory pathways
Protein Data Bank (PDB)	Databases with protein three-dimensional (3D) structures
PEDANT	Integrated resources

1.3 Virtual screening of pharmaceutical agents

1.3.1 Structure-based and ligand based virtual screening

Virtual screening (VS) is a computational technique used in lead compounds discovery research. It involves rapid *in silico* screening of large compound libraries of chemical structures in order to identify those compounds that most likely to interact with a therapeutic target, typically a protein receptor or enzyme

[15, 16]. VS has been widely explored for facilitating lead compounds discovery [17-20], identifying agents of desirable pharmacokinetic and toxicological properties profiling of compounds [21, 22]. There are two main categories of screening techniques: structure-based and ligand-based [23]. **Figure 1-4** shows the general procedure used in SBVS and LBVS.

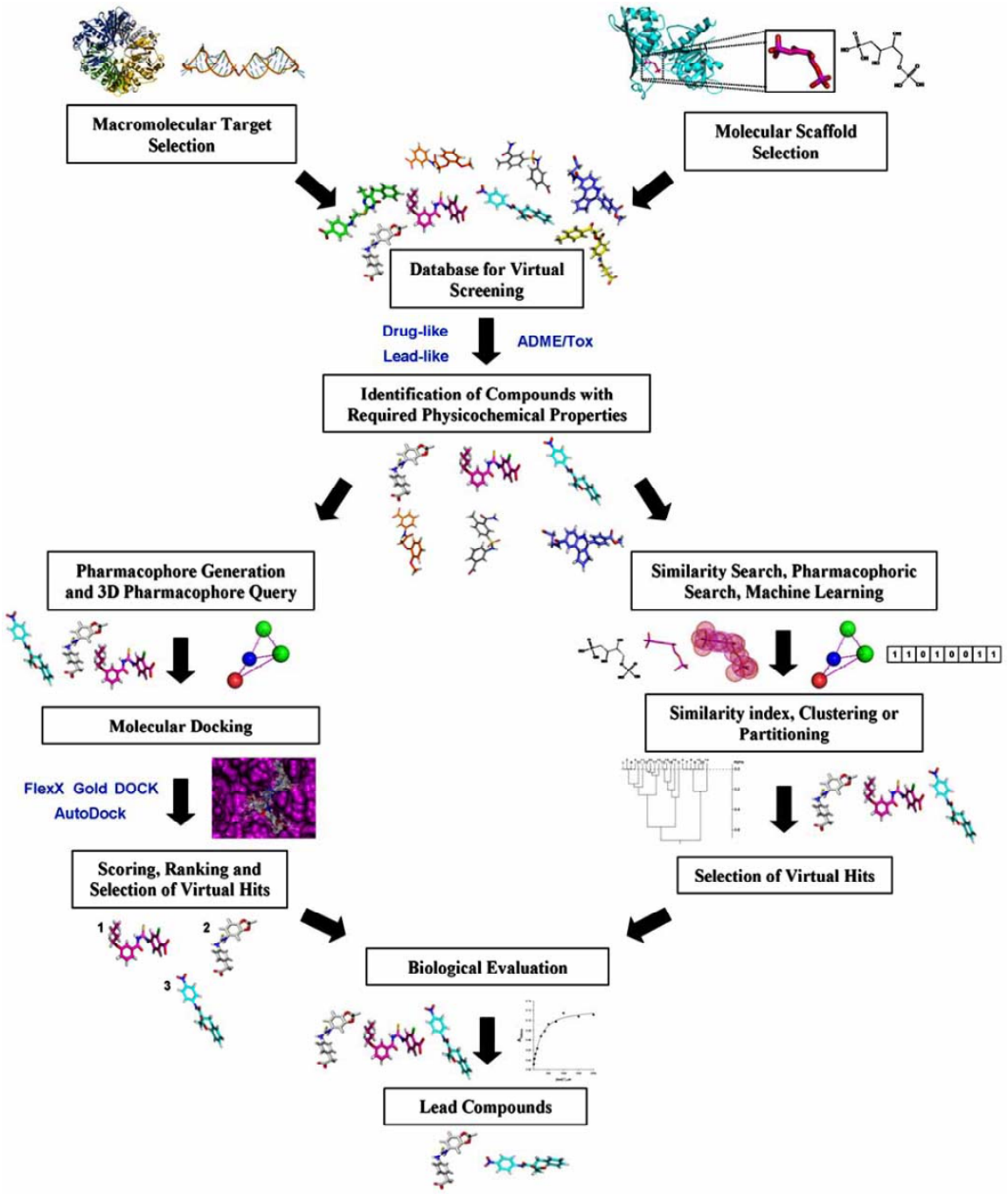


Figure 1-4 General procedure used in SBVS and LBVS (adopted from Rafael V.C. et al[24]).

Structure-based virtual screening (SBVS) begins with a 3-D structure of a target protein and a collection of the 3-D structures of ligands as the screening library. When the 3D structure of a protein target derived either from experimental data (X-ray or NMR spectroscopy) or from homology modeling is available,

SBVS method is applied. SBVS procedure includes docking and scoring. The docking algorithms [25, 26] are designed to evaluate the ligand conformation and orientation within the target surface active site. The scoring methods are empirically or semi-empirically derived to estimate the binding affinities of the ligand and the protein in bound complexes [27]. Docking and scoring algorithms are often merged to detect those compounds with highest affinity against a target by predicting the binding mode (by docking) and affinity (by scoring). So far, more than 60 docking programs and 30 scoring functions have been reported [28, 29]. The major disadvantage of SBVS is the absence of appropriate scoring functions to separate correct and incorrect poses of bound ligands and to identify false negative and positive hits. In addition, the challenges encountered by SBVS include the appropriate treatment of ionization, tautomerization of ligand and protein residues, target/ligand flexibility, choice of force fields, salvation effects, dielectric constants, exploration of multiple binding modes and, most importantly, the approximations in the scoring functions that lead to false-positives and miss true-hits. Moreover, most docking algorithms and scoring functions are tuned towards high throughput, which needs a compromise between the speed and accuracy of binding mode and energy prediction. Despite the successful drug discovery cases, currently there has not been a single docking program that outperforms all others with regard to either docking accuracy or hit enrichment. The hit enrichment is defined as the fraction of true active compounds in, for example, the upper 1% of the ranked VS hit list compared with the average fraction of active compounds in the search space. The performance of a docking program is difficult to evaluate in advance, and depends on the nature and quality of the target structure [28-30]. 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 [29, 31]. The published comparison of docking programs has been critically reviewed [32-34].

Unlike SBVS, Ligand-based virtual screening (LBVS) does not require the protein target 3D structure information. Instead, it takes the structure(s) of one or more active compounds as template(s) to identify a new compound library by chemical and physical properties of the template compound(s). The application of LBVS methods firstly use the digital descriptors of molecular structure, properties, or pharmacophore features and then analyze relationships between the training active compounds and test unknown compounds. Different descriptors are designed to detect connections in molecular physical and chemical properties in order to find new hits. Compared with SBVS, LBVS is computationally efficient and is able to screen very large databases in short time. As a result, the LBVS methods are often applied to sequentially screen large compound libraries before more complex experiments are applied. Many types of LBVS methods have been reported with literally thousands of different descriptors. These descriptors are derived from the 2D or 3D distribution of atomic properties of the known compounds, or from the existence of specific structural elements such as double bonds. Many methods designed for the comparison of the similarity of compounds based on these descriptors. Shape comparison [35] and pharmacophore searches are widely used and long-established techniques [36, 37]. Other methods employ molecular fields to define the similarity of compound structures [38, 39]. When large sets of active and inactive compounds are available, machine learning methods, such as

artificial neural nets, decision trees, support vector machines and Bayesian classifiers, can be used to train predictive VS models that can distinguish active from inactive compounds based on their specific physical and chemical features. Comprehensive reviews of ligand-based VS have been presented in a number of reviews [40, 41]. **Appendix A Tables 2, 3, 4** and **5** provide the comparison of performances of some frequently applied SBVS and LBVS methods for identifying inhibitors, agonists and substrates of proteins of pharmaceutical relevance.

1.3.2 Machine learning methods for virtual screening

With the advancement in computational technologies, machine learning methods have become increasing useful in the drug discovery. Machine learning methods typically include procedures used in the study of computer predictions, classifications or analysis of algorithms where the learning process may improve automatically through experience. In target discoveries, machine learning classification methods have been applied for analyzing microarray data, non-invasive images, and mass spectral data to find biomarkers. In drug lead identification, machine learning classification methods are used for assess potential lead suspects, and for performing ligand based virtual screening to find possible hits. In addition machine learning classification methods are used to eliminate toxic compounds at very early stage of drug discovery.

The most common machine learning methods are support vector machines (SVM), Artificial neural network (ANN), probabilistic neural network (PNN), k nearest

neighbor (K-NN), C4.5 decision tree (C4.5DT), linear discriminate analysis (LDA) and logistic regression (LR) which have shown good performance in various fields. Machine Learning Classification (MLC) methods are increasingly used in early drug discovery stage for targets and leads discovery, prediction of interactions with ABC-transporters [42], early detection of drug-induced idiosyncratic liver toxicity [43], prediction of toxicological properties and adverse drug reactions of pharmaceutical agents [44], prediction of P-glycoprotein substrates [45, 46], prediction of drug-likeness [47-49]. The motivation for the adoption of machine learning classification methods in drug discovery is its capability to model complex relationships in biological data.

Comparing with SBVS and other LBVS methods such as QSAR, pharmacophore and clustering methods [18, 50-56], machine learning methods are more capable of dealing with a more diverse spectrum of compounds and more complex structure-activity relationships. The reason is that machine learning methods apply complex nonlinear mappings from molecular descriptors to activity classes without restriction on structural frameworks, and machine learning method do not require prior knowledge of relevant molecular descriptors and functional form of structure-activity relationships [57-61]. Additionally, machine learning methods can be used to overcome several problems that have obstructed the some conventional virtual screening tools [17, 58], which include the extensiveness and discreteness natures of the chemical space, the absence of protein target structures (only 15% of known proteins have known 3D structures), complexity and flexibility of target structures, limited diversity caused by the biased training molecules, 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 1-4**. These reported studies [62-69] primarily focused on the prediction of compounds that inhibit, antagonize, block, agonize, or activate specific therapeutic target proteins. The majority of the reported screening tasks by machine learning methods are found to demonstrate good performance. The yields, hit rates, and enrichment factors of machine learning methods are in the range of 50%~94%, 10%~98%, and 30~108 respectively. **Table 1-5**, **Table 1-6** and **Table 1-7** show the tentative comparisons of the reported performances of structure-based VS methods and two classes of ligand-based VS methods, pharmacophore and clustering. Most of the yields, hit rates, and enrichment factors lay in the range of 7%~95%, 1%~32%, and 5~1189 for structure-based, 11%~76%, ~0.33%, and 3~41 for pharmacophore, and 20%~63%, 2%~10%, and 6~54 for clustering methods respectively. The general performance of machine learning methods appears to be comparable to or in some cases better than the reported performances of the conventional VS studies such as pharmacophore and clustering methods. In screening extremely-large libraries, the reported yields, hit-rates and enrichment factors of machine learning VS tools are in the range of 55%~81%, 0.2%~0.7% and 110~795 respectively, compared to those of 62%~95%, 0.65%~35% and 20~1,200 by structure-based VS tools. The reported hit-rates of some machine learning VS tools are comparable to those of structure-based VS tools in screening libraries of ~98,000 compounds, but their enrichment factors are substantially smaller. Therefore, while exhibiting equally good yield, in screening extremely-large (≥1 million) and large (130,000~400,000)

libraries, the currently developed machine learning VS tools appear to show lower hit-rates and, in some cases, lower enrichment factors than the best performing structure-based VS tools.

1.3.3 Virtual screening for subtype-selective pharmaceutic agents

Drugs that selectively modulate protein subtypes are highly useful for achieving therapeutic efficacies at reduced side effects [90-93]. For some targets such as dopamine receptors, all of the approved drugs are subtype non-selective, and this non-selectivity directly contributes to their observed side effects and adversely affects their application potential [93]. There is a need for developing subtype selective drugs against these targets [92-96].

Several multi-label machine learning methods have been used for developing *in-silico* tools to predict protein selective compounds within a protein family or subfamily. For instance, multi-label support vector machines (ML-SVM), multi-label k-nearest-neighbor (ML-kNN) and multi-label counter-propagation neural network (ML-CPNN) methods have been used for predicting isoform specificity of P450 substrates [97, 98]. Combinatorial support vector machines (Combi-SVM) method has been used for identifying dual kinase inhibitors selective against single kinase inhibitors of the same kinase pair and inhibitors of other kinases [99].

Consequently, although these methods have shown good performance in selecting ligands of a subtype, they do not always distinguish subtype selective and non-selective ligands at good accuracy levels. For instance, the ML-SVM, ML-kNN and ML-CPNN methods predict 88%, 64% and 34% isoform selective substrates as selective respectively, 99%, 82% and 72% isoform non-selective

substrates as non-selective respectively [97]. Combi-SVM identifies 51.9%-96.3% single kinase inhibitors as kinase selective with respect to a specific kinase pair and 12.2%-57.3% dual kinase inhibitors as dual inhibitors [99]. Therefore, new methods need to be explored for better distinguishing subtype selective and non-selective ligands.

1.4 Bioinformatics tools in biomarker identification

With the advances of biotechnology, the development of molecular biomarkers of exposure, toxicity, disease risk, disease status and response to therapy have been greatly accelerated. A biomarker is a characteristic that is objectively measure and evaluated as an indicators of normal biologic processes, pathogenic processes or pharmacological responses to therapeutic or other interventions[100]. Biomarker studies are aiming to develop a biomarker classifier that can be utilized for disease diagnostics, safety assessment, prognostics and prediction of response for patient treatments [101, 102]. Microarray technology, which is capable of providing the expression profile information on thousands of genes simultaneously, has become a very important component of disease molecular differentiation. The gene expression profiles can be applied to identify markers which are closely associated with early detection/differentiation of disease, or disease behavior (disease progression, response to therapy), and could serve as disease targets for drug design [103]. This strategy is widely used in cancer research for the identification of cancer markers, and provides new insights into tumorigenesis, tumor progression and invasiveness [101, 104-108].

The statistical methods in microarray data analysis can be classified into two

groups: unsupervised learning methods and supervised learning methods. Unsupervised analysis of microarray data aims to group relative genes without knowledge of the clinical features of each sample [109]. A commonly used unsupervised method is hierarchical clustering method. This method groups genes together on the basis of shared expression similarity across different conditions, under the assumption that genes are likely to share the same function if they exhibit similar expression profiles [110-113]. Hierarchical clustering creates phylogenetics trees to reflect higher-order relationship between genes with similar expression patterns by either merging smaller clusters into larger ones, or by splitting larger clusters into smaller ones. A dendogram is constructed, in which the branch lengths among genes also reflect the degree of similarity of expression [114, 115]. Unsupervised methods have some merits such as good implementations available online and the possibility of obtaining biological meaningful results, but they also possess some limitations. First, unsupervised methods require no prior knowledge and are based on the understanding of the whole data set, making the clusters difficult to be maintained and analyzed. Second, genes are grouped based on the similarity that can be affected by input data with poor similarity measures. Third, some of the unsupervised methods require the predefinition of one or more user-defined parameters that are hard to be estimated (e.g. the number of clusters). Changing these parameters often have a strong impact on the final results [116].

In contrast to the unsupervised methods, supervised methods require a priori knowledge of the samples. Supervised methods generate a signature that contains genes associated with the clinical response variable. The number of significant genes is determined by the choice of significance level. SVM [117] and ANN [118]

are two important supervised methods. Both methods can be trained to recognize and characterize complex pattern by adjusting the parameters of the models fitting the data by a process of error (for example, miss-classification) minimization through learning from experience (using training samples). SVM separates one class from the other in a set of binary training data with the hyperplane that is maximally distant from the training examples. This method has been used to rank the genes according to their contribution to defining the decision hyperplane, which is according to their importance in classifying the samples. Ramaswamy et al. used this method to identify genes related to multiple common adult malignancies [105]. ANN consists of a set of layers of perceptrons to model the structure and behavior of neutrons in the human brain. ANN ranks the genes according to how sensitive the output is with respect to each gene's expression level. Khan et al identified genes expressed in rhabdomyosarcoma from such strategy [106].

No matter whether the supervised or unsupervised methods are used, one critical problem encountered in both methods is feature selection, which has become a crucial challenge of microarray data analysis. The challenge comes from the presence of thousands of genes and only a few dozens of samples in currently available data. Therefore, there is a need of robust techniques capable of selecting the subsets of genes relevant to a particular problem from the entire set of microarray data both for the disease classification and for the disease target discovery. Many gene selection methods have been developed, and generally fall into two categories: filter method and wrapper method [119]. In brief, the filter method selects genes independent of the learning algorithms [120-122]. It evaluates the goodness of the genes from simple statistics computed from the

empirical distribution with the class label [123]. Wrapper method generates genes from the evaluation of a learning algorithm. It is conducted in the space of genes, evaluating the goodness of each gene or gene subsets by such criteria as cross-validation error rate or accuracy from the validation dataset [124]. Recursive feature elimination (RFE) is a good example of the wrapper method for disease gene discovery. The RFE method uses the prediction accuracy from SVM to determine the goodness of a selected subset. Machine learning methods such as SVM-RFE are widely used in analyzing microarray data in order to identify biomarkers. However, there are two fundamental problems: One problem is to specify the number of genes for differentiating disease and prognosis of patients. Another problem in gene discovery is the gene signatures were highly unstable and strongly depended on the selection of patients in the training sets. We explore a new signature selection method aiming at reducing the chances of erroneous elimination of predictor-genes due to the noises contained in microarray dataset. Multiple random sampling and gene-ranking consistency evaluation procedures will be incorporated into RFE signature selection method. The consistent genes obtained from the multiple random sampling method may give us a better understanding to the disease initiation, progress and response to treatment.

1.5 Objectives and outline

Overall, there are three objectives for this work:

1. To develop a novel virtual screening method for prediction of subtype selective pharmaceutical agents.
2. To test subtype selective virtual screening model on prediction of selective

ligands of dopamine receptor and to compare with other conventional methods.

3. To develop machine learning based virtual screening method to prediction potential IKK beta inhibitors. In addition, to compare the virtual screening performances of machine learning methods SVM, k-NN and PNN.
4. To identify biomarker for predicting response to anticancer EGFR tyrosine kinase inhibitors.

Target selective drugs are developed for enhanced therapeutics and reduced side effects. *In-silico* methods such as machine learning methods have been explored for searching target selective ligands such as dopamine receptor ligands, but encountered difficulties associated with high subtype similarity and ligand structural diversity. The first aim of thesis is to develop a novel virtual screening method for prediction of subtype selective pharmaceutical agents. We tested the novel method on dopamine receptor subtype selective ligands VS.

Protein Kinases are important regulators of cell function that constitute one of the largest and most functionally diverse gene families. Despite the hundreds of kinase inhibitors currently in discovery and pre-clinical phases, the number of kinase inhibitors drugs that have been approved remains low by comparison. Moreover, some drugs such as anticancer EGFR tyrosine kinase inhibitors elicit markedly different clinical response rates due to differences in drug bypass signaling as well as genetic variations of drug target and downstream drug-resistant genes. In the thesis, we also aimed to develop VS method for facilitating IKK beta inhibitors discovery. In addition, we aimed to identify biomarker for predicting response to anticancer EGFR tyrosine kinase inhibitors by systematically analysis bypassing signaling pathways.

This thesis is outlined as follows:

Chapter 1, an introduction to cheminformatics and bioinformatics is given followed by introduction of virtual screening methods.

Chapter 2 describes methods used in this work, including data collection, machine learning methods, and virtual screening model validation and performance measurements. Finally, techniques for identifying biomarkers by implementing feature reduction algorithm are described.

Chapter 3 shows the development of a novel support vector machines approach for virtual screening of dopamine receptor subtype-selective ligands. Comparison of the performance with multi-label and combinatorial SVM method is also described in this chapter.

Chapter 4 is devoted to the use of virtual screening approach in prediction of IKK beta inhibitors. SVM based VS model is compared with KNN and PNN based VS model in screening large libraries.

Chapter 5 elaborates the analysis of bypass signaling in EGFR pathway and profiling of bypass genes for predicting response to anticancer EGFR tyrosine kinase inhibitors.

In the end, chapter 6 summarizes overall findings of this work and discusses the limitations and suggestions for future study.

Chapter 2 Methods

This chapter includes methods of virtual screening: (1) Datasets, including data collection and quality analysis (section 2.1); (2) Molecular descriptors calculation (section 2.2); (3) Statistical machine learning methods in ligand based virtual screening (section 2.3); (4) Statistical machine learning methods model evaluations (section 2.4); Moreover, feature reduction methods in biomarker identification are also described (section 2.5).

2.1 Datasets

2.1.1 Data Collection

Sufficient and high quality data is critical for drug discovery and especially essential for *in-silico* approaches since they rely on the quantity and quality of the available data. Massive amount of data about small molecules and their related annotation information have been accumulated in scientific literatures and cheminiformatics databases. **Table 2-1** lists some of the widely known small molecule databases.

The datasets used in this work mainly are retrieved from the following two types of sources. First, we collected small molecular data from credible journals such as Bioorganic & Medicinal Chemistry Letters, Bioorganic & Medicinal Chemistry, European Journal of Medicinal Chemistry, European Journal of Organic Chemistry and Journal of Medicinal Chemistry, etc. Second, we use

cheminformatics databases that contain accurate and reliable data such as PubChem and ChEMBL [125].

Table 2-1 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 Quality analysis

The reliability of *in silico* approaches of pharmacological properties classification depends on the availability of high quality pharmacological data with low experimental errors [126]. Ideally, the measurements of pharmacological data properties should be conducted with the same protocol so that there is a common ground to compare different compounds with each other. However, some

pharmacological properties measurements have been used only for a limited number of compounds and most pharmacological properties measurements are rarely determined by the same protocol. Thus the collected data consist of compound data measured by different protocols and the incorporation of additional experimental information. To maintain the data quality, in this work, several methods are adopted to ensure that inter-laboratory variations caused by different experimental protocols do not significantly affect the quality of the training sets. The pharmacological property measurements for data were investigated and the ones that contain large variations in experimental protocols compared to the majority of the data are filtered. It is estimated that the most common range of the pharmacological properties measurements for the compounds investigated in more than one source was used to select compounds for the different classes [127].

Diversity Index (DI) is employed to evaluate the structural diversity of a collection of compounds. It is defined as the average value of the similarity between pairs of compounds in a dataset [128],

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

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 [129] were used to compute *sim(i,j)* in this study,

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

where l is the number of descriptors calculated for the compounds in the datasets, and x is the calculated descriptors. The mean maximum tanimoto coefficient of the compounds in dataset A and those in dataset B can be used as a representative index (RI) to measure the extent to which dataset B is representative of dataset A. Dataset B is more representative of dataset A if the RI value between dataset A and B is higher.

2.2 Molecular descriptors

2.2.1 Definition and generation of molecular descriptors

Molecular descriptors have been extensively used in deriving structure-activity relationships [130, 131], quantitative structure activity relationships [132, 133], and machine learning prediction models for pharmaceutical agents [49, 134-140]. 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 an useful number or the result of some standardized experiment".

Many programs e.g. PaDEL-descriptor [141], DRAGON [142], Molconn-Z [143], MODEL [144], Chemistry Development Kit (CDK) [145, 146], JOELib [147] and Xue descriptor set [138], are available to calculate physical and chemical descriptors. These methods can be applied to derive >3,000 molecular descriptors

[138, 141-147]. These descriptors include constitutional descriptors, topological descriptors, RDF descriptors [148], molecular walk counts [149], 3D-MoRSE descriptors [150], BCUT descriptors [151], WHIM descriptors [152], Galvez topological charge indices and charge descriptors [153], GETAWAY descriptors [154], 2D autocorrelations, functional groups, atom-centred descriptors, aromaticity indices [155], Randic molecular profiles [156], electrotopological state descriptors [157], linear solvation energy relationship descriptors [158], and other empirical and molecular properties. However, not all of the available descriptors are needed to fully represent the features of a particular class of compounds. Contrarily, without appropriate descriptors, the performance of a developed ML VS tool may be affected to some degrees. This is caused by the noise arising from the high redundancy and overlapping of the available descriptors. For example, the Xue descriptor set and 98 1D and 2D descriptors are widely used in machine learning based virtual screening models. These 98 descriptors were selected from the descriptors derived from MODEL program by discarding those that were redundant and unrelated to the problem studied here. The Xue descriptor set and these 98 descriptors are listed in **Table 2-2** and **Table 2-3**.

Table 2-2 Xue descriptor set

Descriptor Class	**Number of descriptor in class**	**Descriptors**
Simple molecular properties	18	Molecular weight, Number of rings, rotatable bonds, H-bond donors, and H-bond acceptors, Element counts
Molecular connectivity and shape	28	Molecular connectivity indices, Valence molecular connectivity indices, Molecular shape Kappa indices, Kappa alpha indices, flexibility index
Electro-topological state	97	Electrotopological state indices, and Atom type electrotopological state indices, Weiner Index, Centric Index, Altenburg Index, Balaban Index, Harary Number, Schultz Index, PetitJohn R2 Index, PetitJohn D2 Index, Mean Distance Index, PetitJohn I2 Index, Information Weiner, Balaban RMSD Index, Graph Distance Index
Quantum chemical properties	31	Polarizability index, Hydrogen bond acceptor basicity (covalent HBAB), Hydrogen bond donor acidity (covalent HBDA), Molecular dipole moment, Absolute hardness, Softness, Ionization potential, Electron affinity, Chemical potential, Electronegativity index, Electrophilicity index, Most positive charge on H, C, N, O atoms, Most negative charge on H, C, N, O atoms, Most positive and negative charge in a molecule, Sum of squares of charges on H,C,N,O and all atoms, Mean of positive charges, Mean of negative charges, Mean absolute charge, Relative positive charge, Relative negative charge
Geometrical properties	25	Length vectors (longest distance, longest third atom, 4th atom), Molecular van der Waals volume, Solvent accessible surface area, Molecular surface area, van der Waals surface area, Polar molecular surface area, Sum of solvent accessible surface areas of positively charged atoms, Sum of solvent accessible surface areas of negatively charged atoms, Sum of charge weighted solvent accessible surface areas of positively charged atoms, Sum of charge weighted solvent accessible surface areas of negatively charged atoms, Sum of van der Waals surface areas of positively charged atoms, Sum of van der Waals

		surface areas of negatively charged atoms, Sum of charge weighted van der Waals surface areas of positively charged atoms, Sum of charge weighted van der Waals surface areas of negatively charged atoms, Molecular rugosity, Molecular globularity, Hydrophilic region, Hydrophobic region, Capacity factor, Hydrophilic-Hydrophobic balance, Hydrophilic 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 this work, the 2D structure of each of the compounds was generated by using ChemDraw or downloaded from databases such as PubChem and BindingDB [159]. Then they were subsequently converted into 3D structure by using CORINA [160]. The 3D structure of each compound was manually inspected to ensure the proper chirality of each chiral agent. All salts and elements, such as sodium or calcium, were removed prior to descriptor calculation. The optimization of generated geometries was conducted without symmetry restrictions. The 3D

structures of the compounds then were used to compute the molecular descriptors by the in-house programs and scripts.

2.2.2 Scaling of molecular descriptors

Chemical descriptors are normally scaled before they can be employed for machine learning. Scaling of chemical descriptors ensures that each of descriptor has unbiased contribution in creating the prediction models[161]. Scaling can be done by a number of ways e.g auto-scaling, range scaling, Pareto scaling, and feature weighting [162, 163]. In this work, range scaling is used to scale the chemical descriptor data. Range scaling is done by dividing the difference between descriptor value and the minimum value of that descriptor with the range of that descriptor:

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

Where d_{ij}^{scaled}, d_{ij} $_{ij}$, $d_{j,\max}$ 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 Statistical machine learning methods in ligand based virtual screening

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. Websites for the freely downloadable codes of some methods are given in **Table 2-4**.

Table 2-4 Websites that contain freely downloadable codes of machine learning methods.

BKD	
Binding Database	http://www.bindingdb.org/bind/vsOverview.jsp
Decision Tree	
PrecisionTree	http://www.palisade.com.au/precisiontree/
DecisionPro	http://www.vanguardsw.com/decisionpro/jdtree.htm
C4.5	http://www2.cs.uregina.ca/~hamilton/courses/831/notes/ml/dtrees/c4.5/tutorial.html
C5.0	http://www.rulequest.com/download.html
KNN	
k Nearest Neighbor	http://www.cs.cmu.edu/~zhuxj/courseproject/knndemo/KNN.html
PERL Module for KNN	http://aspn.activestate.com/ASPN/CodeDoc/AI-Categorize/AI/Categorize/kNN.html
Java class for KNN	http://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/classify/old/KNN.html
LDA	
DTREG	http://www.dtreg.com/lda.htm
LR	
Paul Komarek's Logistic Regression Software	http://komarix.org/ac/lr/lrtrirls
Web-based logistic regression calculator	http://statpages.org/logistic.html
Neural Network	
BrainMaker	http://www.calsci.com/
Libneural	http://pcrochat.online.fr/webus/tutorial/BPN_tutorial7.html
fann	http://leenissen.dk/fann/
NeuralWorks Predict	http://www.neuralware.com/products.jsp
NeuroShell Predictor	http://www.mbaware.com/neurpred.html
SVM	
SVM light	http://svmlight.joachims.org/
LIBSVM	http://www.csie.ntu.edu.tw/~cjlin/libsvm/
mySVM	http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html
BSVM	http://www.csie.ntu.edu.tw/~cjlin/bsvm/
SVMTorch	http://www.idiap.ch/learning/SVMTorch.html

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[164, 165], which consistently shows outstanding classification performance, is less penalized by sample redundancy, and has lower risk for over-fitting[166, 167]. 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:

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

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

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(**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 compounds of diverse structures [168-175], 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

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 [176-178]. 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. For the SVM model in this study, hard margin SVM was used and σ was scanned between 0 and 15 for the best performing performing model. Software LibSVM[179], an integrated software for support vector classification, regression and distribution estimation, was chosen to do the machine learning in this work.

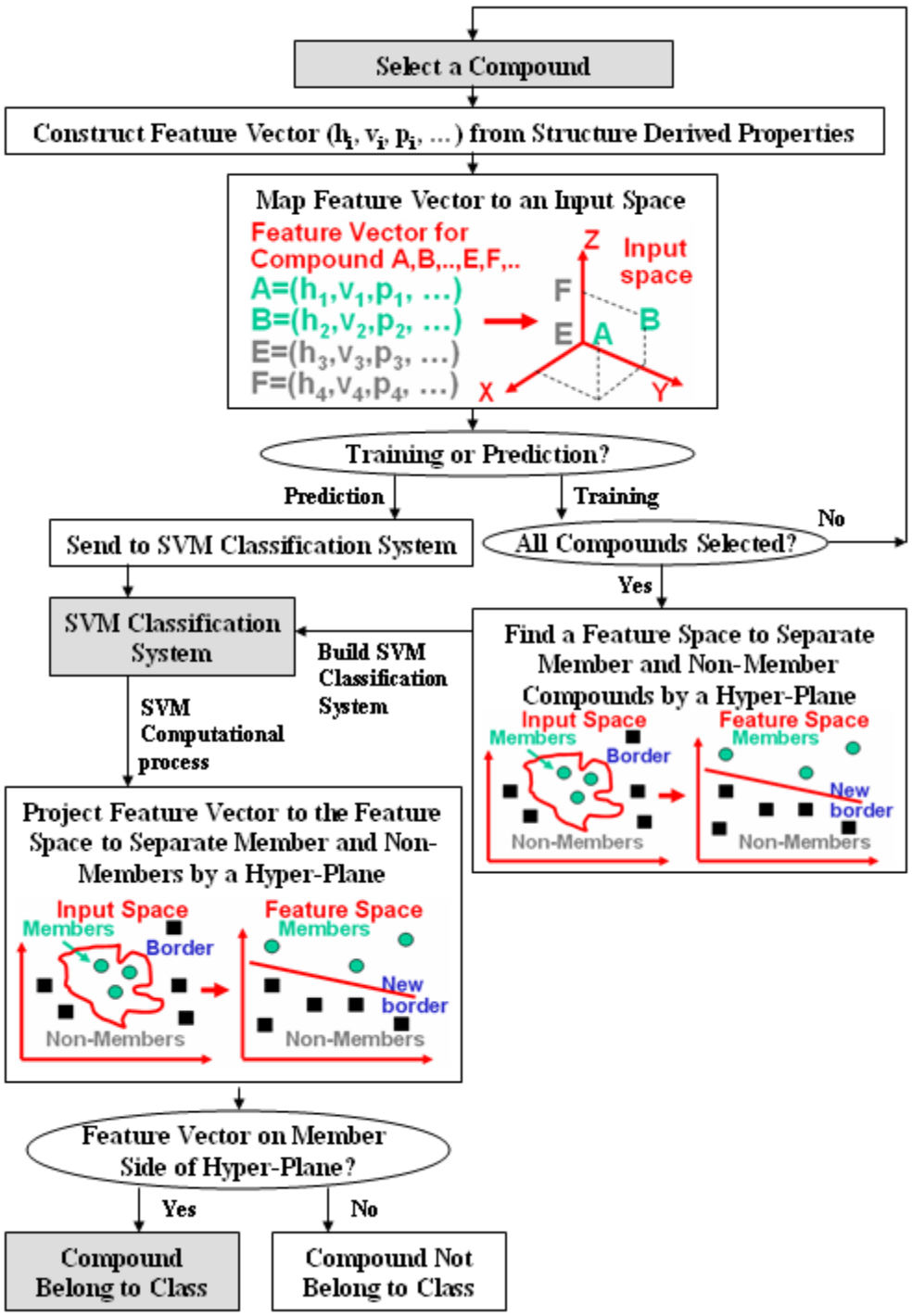


Figure 2-1 Schematic diagram illustrating the process of the training a prediction model and using it for predicting active compounds of a compound class from their structurally-derived properties (molecular descriptors) by using support vector machines. A, B, E, F and (h_j, p_j, vj,…) represents such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

2.3.2 K-nearest neighbor method

K-NN measures the Euclidean distance $D = \sqrt{\|\mathbf{x} - \mathbf{x}_i\|^2}$ between a compound ***x*** and each individual inhibitor or non-inhibitor $\mathbf{x}_i$ in the training set[180, 181]. A total of *k* number of vectors nearest to the vector ***x*** are used to determine the decision function *f(**x**)*:

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

where $\delta(a,b) = 1$ if $a = b$ and $\delta(a,b) = 0$ if $a \neq b$, arg max 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 neighbors. The procedure of k-NN is illustrated in **Figure 2-2**.

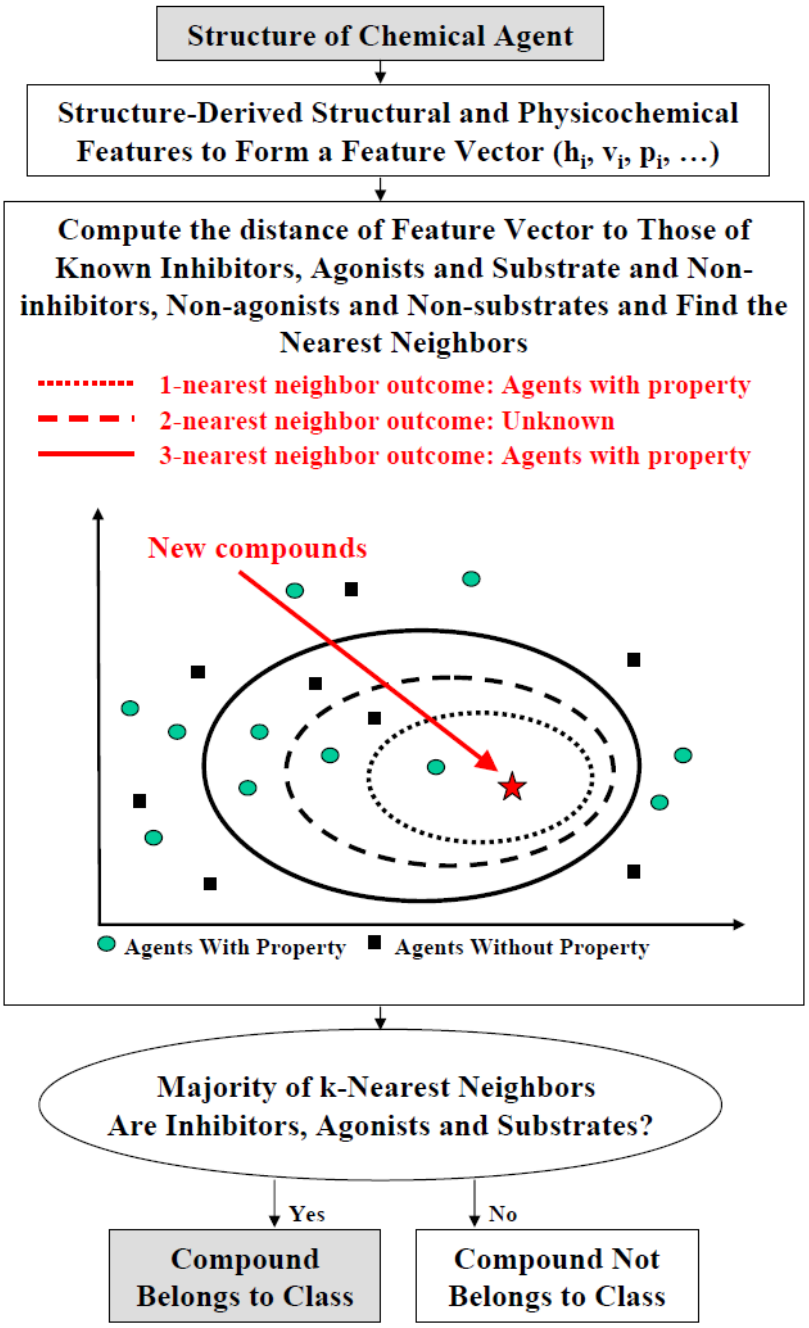


Figure 2-2 Schematic diagram illustrating the process of the prediction of compounds of a particular property from their structure by using a machine learning method – k-nearest neighbors (K-NN). A, B: feature vectors of agents with the property; E, F: feature vectors of agents without the property;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

2.3.3 Probabilistic neural network method

Probabilistic Neural Network (PNN) belongs to the neural network methods. It is designed for classification through the use of Bayes' optimal decision rule [127]: $h_i c_i f_i(\mathbf{x}) > h_j c_j f_j(\mathbf{x})$, where h_i and h_j are the prior probabilities, c_i and c_j are the costs of misclassification and $f_i(x)$ and $f_j(x)$ are the probability density function for class i and j respectively. An unclassified vector $\boldsymbol{x}$ is classified into population i if the product of all the three terms is greater for class i than for any other class j (not equal to i). In most applications, the prior probabilities and costs of misclassifications are treated as being equal. The probability density function for each class for a univariate case can be estimated by using the Parzen's nonparametric estimator[182],

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

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.

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

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

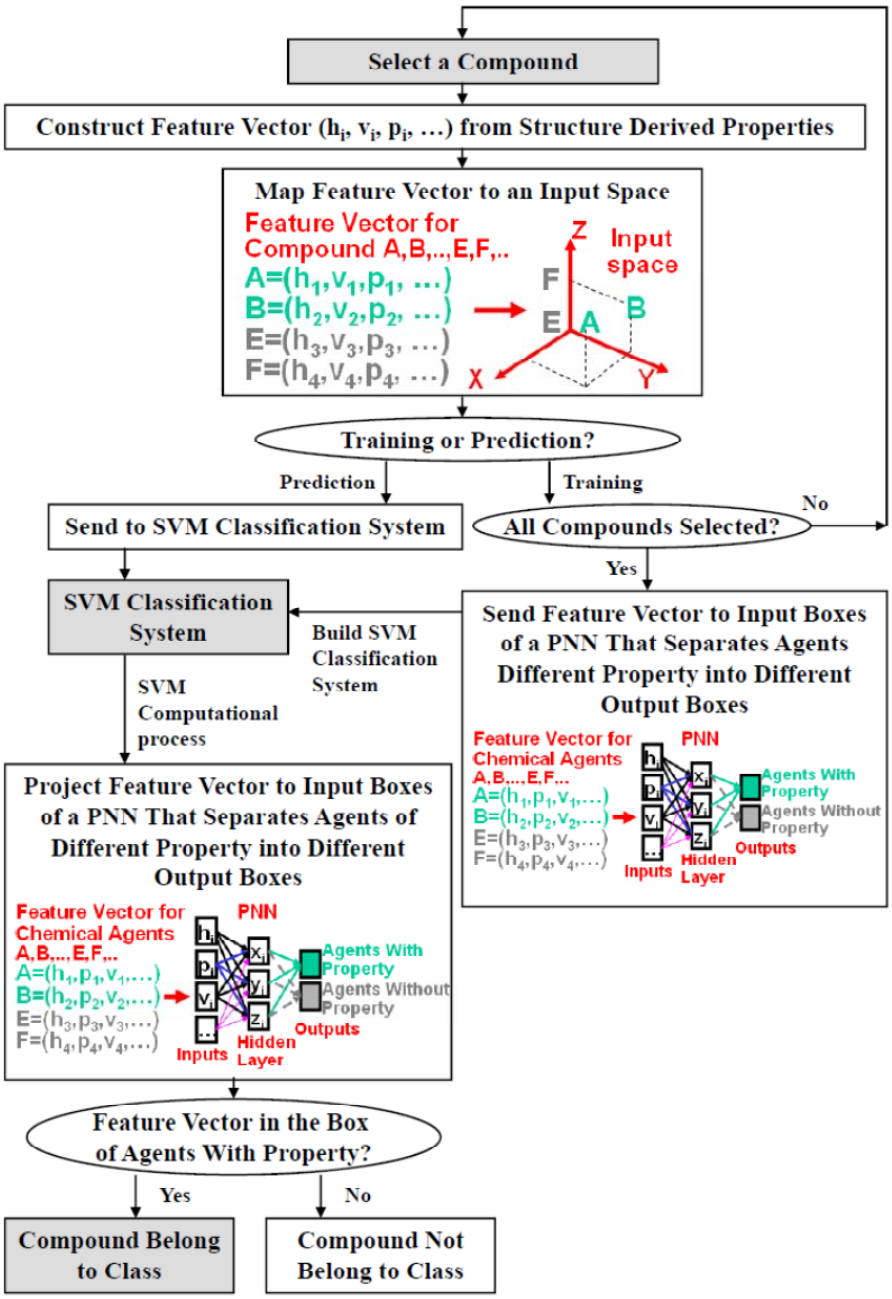


Figure 2-3 Schematic diagram illustrating the process of the prediction of compounds of a particular property from their structure by using a machine learning method –probabilistic neural networks (PNN). A, B: feature vectors of agents with the property; E, F: feature vectors of agents without the property;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

Determining if two compounds are similar to each other or not in a training dataset can be conducted by using the Tanimoto coefficient *sim(i,j)* [129]

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

where *l* is the number of molecular descriptors. A compound *i* is considered to be similar to a known active *j* in the active dataset if the corresponding *sim(i,j)* value is greater than a cut-off value. In this work, in computing *sim(i,j)*, the molecular descriptor vectors $\mathbf{x}_i$s were scaled with respect to all of the MDDR. The cut-off values for similarity compounds are typically in the range of 0.8 to 0.9 [183, 184].

2.3.5 Combinatorial SVM method

In combinatorial strategy, SVM models for each receptor subtype are separately constructed, which are subsequently used for parallel screening against each individual subtype to find compounds that only bind to one of the subtypes (putative subtype selective ligands) or simultaneously bind to multiple subtypes (putative subtype non-selective ligands) [99, 232]. **Figure 2-4** shows the process of combinatorial SVM method.

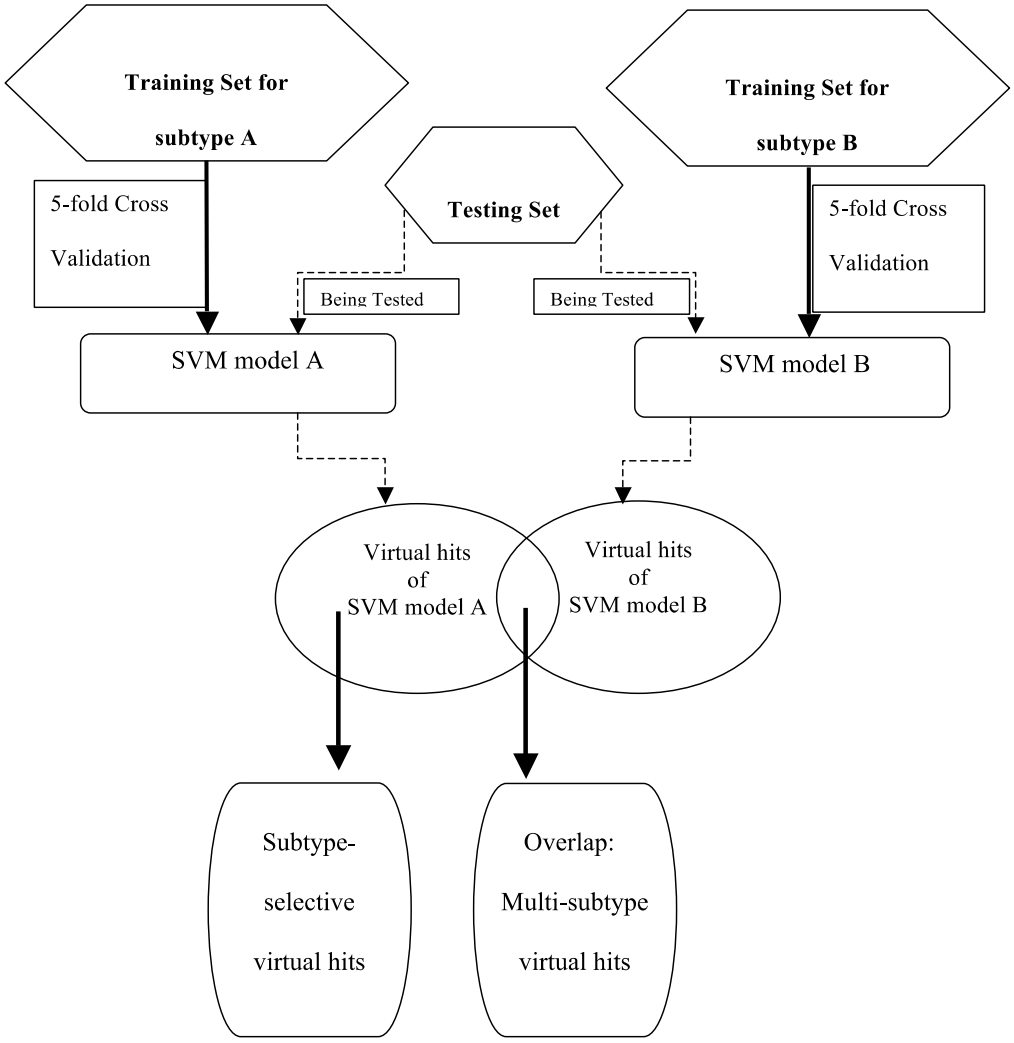


Figure 2-4 Schematic diagram of combinatorial SVM method.

2.3.6 Two-step Binary relevance SVM method

Subtype selective ligands were selected by two steps (**Figure 2-5**). In the first step, a high performance SVM model was developed for each receptor subtype to select ligands of that subtype regardless of their selectivity towards other subtypes. The high performance in selecting ligands of a subtype was achieved by using comprehensive sets of known ligands and putative non-ligands of the corresponding receptor to train the respective SVM model [206]. In the second step, the Binary relevance (BR) method [215] was used for more refined selection

of subtype selective ligands from the putative ligands selected in the first step. BR is a popular multiple binary classification method that transforms the original N-label dataset into N pairs of datasets with samples of each label as positive dataset and samples of the other N-1 labels as negative dataset [215].

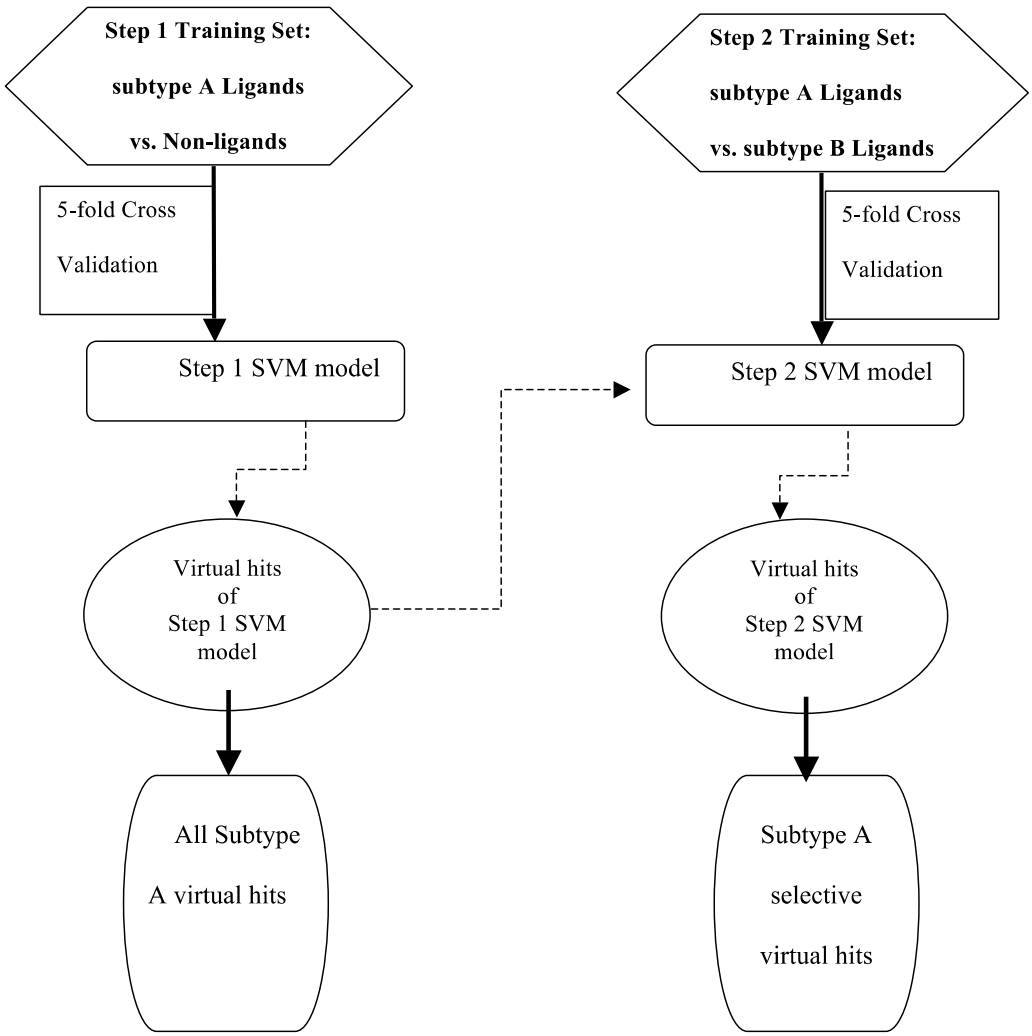


Figure 2-5 Schematic diagram of two-step binary relevance SVM method.

2.4 Statistical machine learning methods model evaluations

2.4.1 Model validation and parameters optimization

Different Statistical learning methods (SLMs) have types of parameters that must

be optimized. In this work SVM is trained by using a Gaussian radian basis kernel function that has an adjustable parameter σ. For PNN, the only parameter to be optimized is a scaling parameter σ. In KNN, the optimum number of nearest neighbors, *k*, needs to be derived for each training set. Optimization of the parameter for each of these SLMs is conducted by scanning the parameter through a range of values. The set of parameters that produces the best pharmacological property prediction model, which is determined by using cross-validation methods, such as 5-fold cross-validation, 10-fold cross-validation or a modeling testing set, is used to construct a final prediction model which is then further validated to ensure that it is valid and useful for further prediction. One of the usual ways to assess or to find the optimum parameters for a model built by machine learning is to see its performance either by independent validation set or cross validation. In this work, models were validated by using both manually segregated a part of data as independent validation set, and also by cross validation. There are various types of cross validation commonly used in many statistical studies such as repeated random sub-sampling cross validation, k-fold cross validation, and leave one out cross validation. In this work, we have applied 5-fold cross validation. For 5-fold cross-validation, these compounds are randomly divided into five subsets of equal size. Each of these folds contains equal number of positive and negative data, thereby rendering it a stratified cross-validation. Four subsets are selected as the training set and the fifth as the validation set. This process is repeated five times such that every subset is selected as a validation set once. The SVM models were saved in each case and prediction was done for validation data.

2.4.2 Performance evaluation methods

Measurements such as sensitivity, specificity and the overall prediction accuracy are employed to quantitatively assess the performance of virtual screening models. They are defined in terms of true positives TP (pharmaceutical agents possessing a specific pharmacological property), true negatives TN (pharmaceutical agents not possessing a specific pharmacological property), false positives FP (pharmaceutical agents not possessing a specific pharmacological property but predicted as agents possessing the specific pharmacological property) and false negatives FN (pharmaceutical agents possessing a specific pharmacological property but predicted as agents not possessing the specific pharmacological property). Sensitivity and specificity are the measurement of prediction accuracy for pharmaceutical agents possessing a specific pharmacological property and agents not possessing that pharmacological property respectively. The overall prediction accuracy (Q) and Matthews correlation coefficient (MCC) [185] are used to measure the overall prediction performance. SE, SP, Q and MCC are defined as follows:

$$SE = \frac{TP}{TP + FN} \tag{11}$$

$$SP = \frac{TN}{TN + FP} \tag{12}$$

$$Q = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

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

The typical measurements of a model performance in screening large libraries include [58] 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):

$$\text{Yield} = SE \tag{15}$$

$$\text{Hit-rate} = TP/(TP+FP) \tag{16}$$

$$\text{False hit-rate} = FP/(TP+FP) \tag{17}$$

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

2.4.3 Overfiting

Overfitting is a major concern in machine learning classification methods. It happens when a model that agrees well with the observed data but has no predictive ability, which means it does not have any value to unseen or future data. There are two main types of overfitting situations: (1) a model more flexible than it needs to be and (2) a model including irrelevant descriptors [186]. An over-fitted classification system tends to obtain much higher prediction accuracies in the cross-validation sets than in the independent validation sets. Hence frequently used method for checking whether a model is overfitted is to compare the prediction accuracies in the cross-validation procedure with those found in testing independent validation sets [186].

2.5 Feature reduction methods in biomarker identification

2.5.1 Data normalization

The purpose of normalization is to remove systematic variations from the expression values, so that biological difference can be easily distinguished and the comparison of expression levels across samples can be performed. In microarray experiments, all the values are fluorescent intensities, which are directly comparable. Therefore the normalization among genes and arrays [187] are both possible.

The popular normalization methods for microarray experiments include global normalization using all genes on the array, and housekeeping genes normalization using constantly expressed housekeeping/invariant genes [188]. Since housekeeping genes are not as constantly expressed as assumed previously [189], using housekeeping genes normalization might introduce extra potential sources of error. It was further approved that normalization using a reduced subset of genes was less statistically robust than the normalization using the entire gene set [190]. Currently, a typical normalization procedure is (1) normalizing the expression levels of each sample to zero-mean and unit variance, and then (2) normalizing the expression levels of each gene to zero-mean and unit variance over all the samples. This normalization method has been shown to perform well [191, 192] and is applied in this project.

2.5.2 Recursive features elimination SVM

a. Overview of the gene selection procedure

A novel gene selection procedure method based on Support Vector Machines

classifier, recursive feature elimination, multiple random sampling strategies and multi-step evaluation of gene-ranking consistency was established (**Figure 2-4**):

(1) After preprocessing the original data, by using random sampling method, a large number of training-test sample combinations are generated from the original data set.

(2) The large number of sample combinations is divided into *n* groups, and each group contains 500 sample combinations.

(3) In each training-test sample combination of each group, SVM and RFE are used to classify the samples (SVM classifiers) and rank the genes (RFE gene rank criteria). Therefore 500 gene ranking sequences are formed.

(4) The consistency evaluation can be performed based on the 500 sequences and a certain number of genes (for example, *k* genes) can be eliminated.

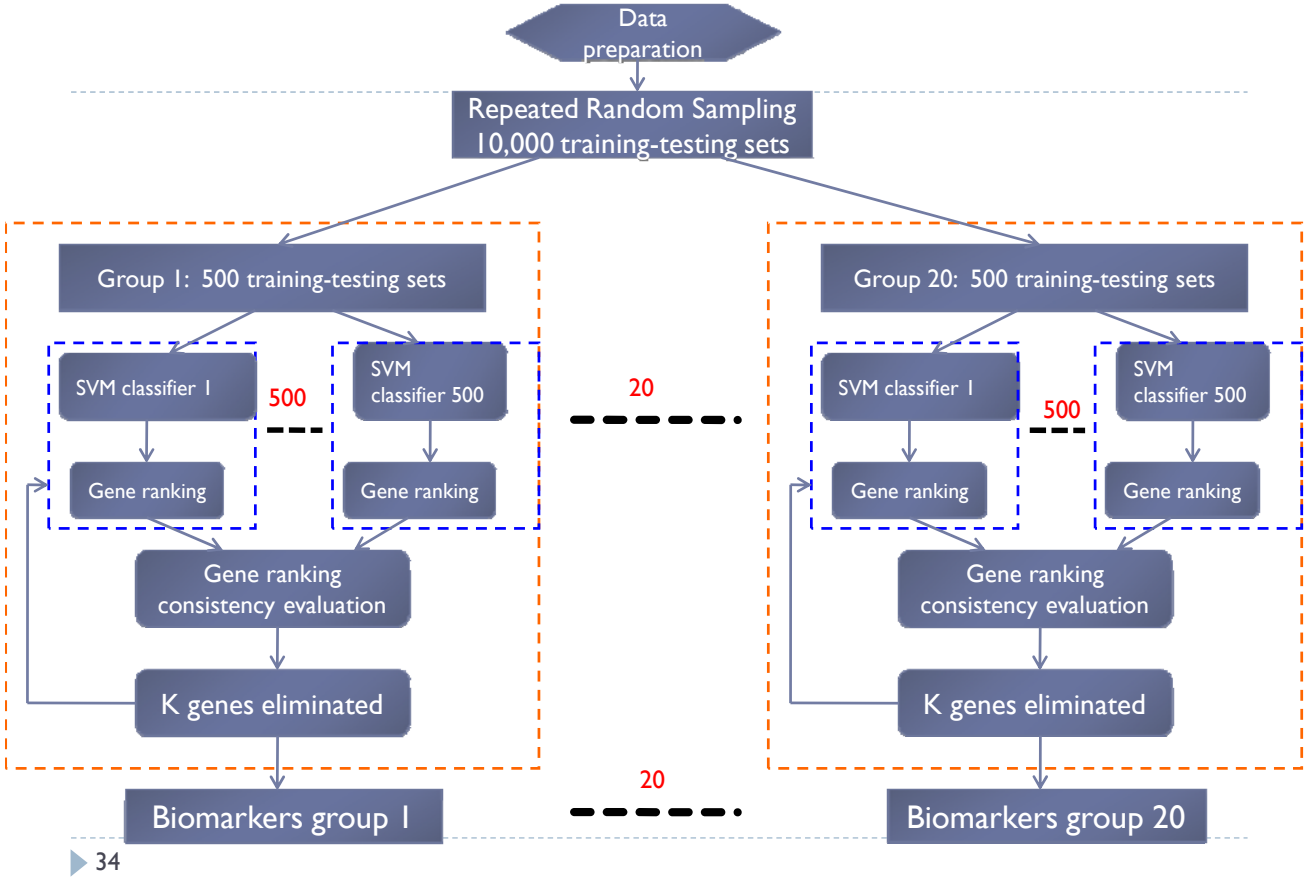
(5) Step (3) and (4) can be iteratively done until no gene can be eliminated.

(6) The gene subset, which gives us the highest overall accuracies of the 500 test sample sets, can be selected as gene signatures of this group. By this way, we can obtain *n* gene signatures.

(7) The stability evaluation of the gene signatures can be performed by looking into the overlap gene rate of the *n* gene signatures.

Below Recursive feature elimination is introduced first and followed by a detailed introduction of the whole feature selection procedure.

Figure 2-6 Overview of the gene selection procedure.



b. Recursive feature elimination

During gene selection procedure, the genes ranked according to their contribution to the SVM classifiers. The contributions of genes are calculated by Recursive feature elimination (RFE) procedure, which sort genes according to a gene-ranking function generated from SVM classifier. SVM training process needs to find the solution for the optimum problem (also known as objective function or cost function) shown in equation:

$$J = \frac{1}{2}\alpha^T H\alpha - \alpha^T 1$$

Under the constraints $\sum_{i=1}^{n} a_i y_i = 0$ and $\alpha_i \geq 0$, i=1,2,…n.

Where $H(i,j) = y_i y_j K(x_i, x_j)$, K is the kernel function.

The gene-ranking function of RFE can be defined as the change in the objective function J upon removing a certain gene. When a given feature is removed or its weight w_k is reduced to zero, the change in the cost function $J(k)$ is

$$DJ(k) = \frac{1}{2}\frac{\partial^2 J}{\partial w_k^2}(Dw_k)^2$$

where the change in weight $Dw_k = w_k - 0$ corresponds to the removal of feature k.

Under the assumption that the removal of one feature will not significantly influence the values of the change of cost function can be estimated as

$$DJ(k) = \frac{1}{2}\alpha^T H\alpha - \frac{1}{2}\alpha^T H(-k)\alpha$$

Where H is the matrix with elements $y_i y_j K(x_i, x_j)$, and $H(-k)$ is the matrix computed by using the same method as that of matrix H but with its kth component removed.

The change in the cost function indicates the contribution of the feature to the

decision function, and serves as an indicator of gene ranking position [193].

c. Sampling, feature elimination and consistency evaluation

In order to present statistical meaning, gene selection is conducted based on multiple random sampling. Each random sampling divide all microarray samples into a training set which contains half number of samples and an associates test set which contains another half number of samples. This sampling method can be treated as a special case of the bootstrap technique. Many researchers showed that bootstrap-related techniques present more accurate estimation than cross-validation on small sample sets [194, 195]. By using this random sampling, thousands of training-test sets, each containing a unique combination of samples, are generated. These thousands of randomly generated training-test sets are randomly divided into several sampling groups, with equal number of training-test sets (such as 500 traing-test sets) in each group. Every sampling group is then used to derive a signature by RFE-SVM.

In every training-test sampling group generated by multiple random sampling, each training-set (totally 500 training-test sets) is used to train a SVM class-differentiation system and the genes are ranked by using Recursive feature elimination (RFE), according to the contribution of genes to the SVM classifier. In order to derive a gene-ranking criterion consistent for all iterations and all the 500 training-test sets in this group, a SVM class-differentiation system with a universal set of globally optimized parameters, which give the best average class-differentiation accuracy over all of the 500 test sets in this group, is applied by RFE gene-ranking function at every iteration step and for every training-test set.

To further reduce the chance of erroneous elimination of predictor-genes,

additional gene-ranking consistency evaluation steps are implemented on top of the normal RFE procedures in each group:

(1) For every training-set, subsets of genes ranked in the bottom (which give least contribution to the SVM classification procedure) with combined score lower than the first few top-ranked genes (which give highest contribution to the SVM classification procedure) are selected such that collective contribution of these genes less likely outweigh top-ranked ones.

(2) For every training-set, the step (1) selected genes are further evaluated to choose those not ranked in the upper 50% in previous iteration so as to ensure that these genes are consistently ranked lower.

(3) A consensus-scoring scheme is applied to step (2) selected genes such that only those appearing in most of the 500 testing-sets were eliminated.

For each sampling group, different SVM parameters are scanned, various RFE iteration steps are evaluated to identify the globally optimal SVM parameters and RFE iteration steps that give the highest average class-differentiation accuracy for the 500 testing-sets.

Chapter 3 A two-step Target Binding and Selectivity Support Vector Machines Approach for Virtual Screening of Dopamine Receptor Subtype-Selective Ligands

Summary

Target selective drugs, such as dopamine receptor (DR) subtype selective ligands, are developed for enhanced therapeutics and reduced side effects. *In-silico* methods have been explored for searching DR selective ligands, but encountered difficulties associated with high subtype similarity and ligand structural diversity. Machine learning methods have shown promising potential in searching target selective compounds. Their target selective capability can be further enhanced. In this work, we introduced a new two-step support vector machines target-binding and selectivity screening method for searching DR subtype-selective ligands, which was tested together with three previously-used machine learning methods for searching D1, D2, D3 and D4 selective ligands. It correctly identified 50.6%-88.0% of the 21-408 subtype selective and 71.7%-81.0% of the 39-147 multi-subtype ligands. Its subtype selective ligand identification rates are significantly better than, and its multi-subtype ligand identification rates are comparable to the best rates of the previously used methods. Our method produced low false-hit rates in screening 13.56M PubChem, 168,016 MDDR and 657,736 ChEMBLdb compounds. Molecular features important for subtype

selectivity were extracted by using the recursive feature elimination feature selection method. These features are consistent with literature-reported features. Our method showed similar performance in searching estrogen receptor subtype selective ligands. Our study demonstrated the usefulness of the two-step target binding and selectivity screening method in searching subtype selective ligands from large compound libraries.

3.1 Introduction

Drugs that selectively modulate protein subtypes are highly useful for achieving therapeutic efficacies at reduced side effects [90-93]. For some targets such as dopamine receptors, all of the approved drugs are subtype non-selective, and this non-selectivity directly contributes to their observed side effects and adversely affects their application potential [93]. There is a need for developing subtype selective drugs against these targets [92-96].

The drug-binding domains of some protein subtypes are highly similar to each other. For instance, the sequence similarities among the transmembrane regions of dopamine receptor subtypes are at high levels of 72%, 73% and 90% between D2-like subfamily members D2 and D4, D3 and D4, and D2 and D3 respectively [196], and at the levels of 68%, 70% and 66% between dopamin receptor subtypes D1 and D2, D1 and D3 and D1 and D4 respectively. Ligand binding selectivity to these subtypes is both determined by the structural and physicochemical features of the conserved and non-conserved residues [197]. For instance, while D2 receptor and D3 receptor share high sequence identity in the seven helices regions that make up most of the binding sites, different compositions of the loop regions affect the contour and topography of the binding pockets and hydrogen bonding sites, which enables subtype selective binding [198, 199]. On the other hand, D2/D4 selectivity has been suggested to be determined by mutated residues within the second, third, and seventh membrane-spanning segments [197].

The high sequence similarity levels make it more difficult to develop dopamine receptor subtype-selective drugs. Efforts have been made in exploring *in-silico* methods for searching dopamine receptor subtype-selective drug leads against highly similar subtypes. For instance, 3D-QSAR models have been developed for D2, D3 and D4 selective ligands respectively, achieving good prediction performances with R^2 and Q^2 values in the ranges of 0.89-0.97 and 0.58-0.84 respectively [198-201]. A GALAHAD based selective pharmacophore model has been constructed for D1/D2 selective agents [202]. CoMFA and CoMSIA models have been developed for D2, D3 and D4 selective ligands [203].

These models have been developed by using 12-163 ligands. Significantly higher numbers of dopamine receptor ligands including subtype selective [91, 93] and multi-subtype [204, 205] ligands have been reported. These ligands are of high structural diversity. The published D1, D2, D3 and D4 ligands are distributed in 225, 642, 463 and 433 compound families (**Table 3-1**) compared to the 90-388 families covered by the inhibitors of many kinases [99]. These structurally diverse ligands are not expected to be fully presented by the existing models trained from limited numbers of ligands. More extensive exploration of the available ligands is needed for developing more effective *in-silico* tools for searching subtype-selective dopamine receptor ligands.

Table 3-1 Datasets of our collected dopamine receptor D1, D2, D3 and D4 ligands, non-ligands and putative non-ligands. Dopamine receptor D1, D2, D3 and D4 ligands (Ki <1μM) and non-ligands (ki >10μM) were collected as described in method section, and putative non-ligands were generated from representative compounds of compound families with no known ligand. These datasets were used for training and testing the multi-label machine learning models.

Dopamine Receptor Subtype	**Training Dataset**			**Independent Testing Dataset**	
	Positive Samples	**Negative Samples**		**Positive Samples**	**Negative Samples**
	Ligands published before 2010 (No of chemical families covered by ligands)	Non-ligands published before 2010	Putative non-ligands	Ligands published since 2010 (percent of ligands outside training chemical families)	Non-ligands published since 2010
D1	491 (225)	264	65198	59 (25.42%)	25
D2	2202 (642)	1577	63687	135 (16.30%)	65
D3	1355 (463)	631	62927	76 (18.42%)	28
D4	1486 (433)	526	63272	29 (34.48%)	33

Machine learning methods are particularly useful for developing virtual screening (VS) models from structurally diverse compounds and for searching large chemical libraries [206-208]. The purchasable real chemical libraries have been expanded to >1 million compounds [209] and the public chemical databases have been expanded at faster paces with PubChem [210], ZINC [211], and ChEMBL [212] databases accumulating >30 million compounds, >13 million purchasable compounds, and >1 million bioactive compounds respectively. The available chemical space defined by these databases may be more extensively explored by the use of machine learning methods [213, 214].

Moreover, several multi-label machine learning methods have been used for developing *in-silico* tools to predict protein selective compounds within a protein family or subfamily. For instance, multi-label support vector machines

(ML-SVM), multi-label k-nearest-neighbor (ML-kNN) and multi-label counter-propagation neural network (ML-CPNN) methods have been used for predicting isoform specificity of P450 substrates [97, 98]. Combinatorial support vector machines (Combi-SVM) method has been used for identifying dual kinase inhibitors selective against single kinase inhibitors of the same kinase pair and inhibitors of other kinases [99]. It is of interest to explore some of these methods and to evaluate their capability in predicting subtype selective dopamine receptor ligands.

These existing methods are based on statistical learning algorithms trained by compounds active and inactive against a specific protein or subtype [97-99, 206]. In these algorithms, the inactive chemical space can be represented by a large number of inactive compounds in a training dataset that typically include representative compounds of chemical families or biological classes. In particular the inactive training dataset of a subtype is typically too large to further add sufficient number of active compounds of other subtypes [97-99, 206]. Consequently, although these methods have shown good performance in selecting ligands of a subtype, they do not always distinguish subtype selective and non-selective ligands at good accuracy levels. For instance, the ML-SVM, ML-kNN and ML-CPNN methods predict 34%-89% isoform selective substrates as selective and 82%-99% isoform non-selective substrates as non-selective [97]. Combi-SVM identifies 51.9%-96.3% single kinase inhibitors as kinase selective with respect to a specific kinase pair and 12.2%-57.3% dual kinase inhibitors as dual inhibitors [99]. Therefore, new methods need to be explored for better distinguishing subtype selective and non-selective ligands.

In this work, we introduced a new method, the two-step binary relevance SVM (2SBR-SVM) method for improving the ability in distinguishing subtype selective and non-selective ligands. Our method adopts a two-step approach, with the first step focusing on the identification of putative ligands of a receptor subtype regardless of their possible binding to other subtypes, and the second step focusing on the further separation of subtype selective and multi-subtype ligands. In the first step, a SVM model was developed for each receptor subtype to select putative ligands regardless of their possible binding to other subtypes using the same method as that described in our earlier studies [206]. In the second step, the Binary relevance (BR) method [215] was used for more refined separation of subtype selective and multi-subtype ligands. Specifically, the training datasets of the multiple receptor subtypes were re-arranged into multiple new training datasets, one for each subtype. For a particular subtype, the ligands of that subtype were used as positive samples and the ligands of the other subtypes as the negative samples to train a SVM model for maximally separating ligands of a subtype with those of other subtypes.

Our new method 2SBR-SVM was tested together with three previously-used methods Combi-SVM [99] and two methods in the Mulan software package [215]: the ML-kNN [97, 216] and Random k-labelset Decision Tree (RAkEL-DT) [217, 218] methods. The purpose of these tests was to evaluate the performance of the previously used methods, and to determine to what extent our new method can improve the performance in selecting dopamine subtype selective ligands.

A number of dopamine receptor subtype selective ligands have been therapeutically explored. For instance, most currently used dopamine agonists for the symptomatic treatment of Parkinson's disease are selective for D2-like receptors primarily because drugs acting on both the D1 and D2 receptor families tend to additively produce motor complications such as dyskinesia [219]. D2-selective drugs have exhibited therapeutic efficacy in animal studies [220] and clinical trials [221]. D3-selective drugs have been explored for the treatment of schizophrenia and drug addiction [222, 223]. D4-selective ligands have shown therapeutic potential against erectile dysfunction [224, 225]. Efforts have also been directed to the development of D1-selective [226, 227] ligands against Parkinson's disease and other related CNS disorders. Therefore, our tests were conducted on D1, D2, D3 and D4 selective and non-selective ligands.

Our VS models were trained from 491-2202 dopamine receptor D1, D2, D3, and D4 ligands published before 2010 with all the known subtype selective ligands and some known multi-subtype ligands excluded. The reason for the exclusion of these subtype selective and multi-subtype ligands from the training process is to test to what extent our VS models can identify subtype selective ligands without explicit knowledge of the known subtype selective and multi-subtype ligands. The prediction performance of these models was evaluated by 29-135 known D1, D2, D3 and D4 ligands and 25-65 known non-ligands published since 2010 and not in the training datasets. The subtype selectivity of these models was tested on the 21-408 known subtype selective ligands and the 39-147 known multi-subtype ligands not in the training datasets.

The performance of our new method, 2SBR-SVM, and the method developed in our previous studies, Combi-SVM [99], in screening large compound libraries was evaluated by 13.56 million PubChem compounds [210], 168,016 MDL Drug Data Report (MDDR) database compounds, and 657,736 ChEMBLdb compounds [125] which represent general chemical libraries, patented bioactive agents, and published bioactive compounds respectively. The capability of 2SBR-SVM in identifying subtype selective ligands of other receptors was further evaluated against estrogen receptor (ER) ERα and ERβ subtype ligands by using the same training and testing procedures as those of the dopamine receptor subtype ligands.

3.2 Method

3.2.1 Datasets

D1, D2, D3 and D4 ligands and non-ligands were collected from comprehensive search of literatures [223, 226, 228, 229] and ChEMBLdb database [125] by using combinations of keywords: “dopamine”, “D1 receptor”, “D2 receptor”, “D3 receptor”, “D4 receptor”, “ligand”, “binding”, “binder”, “subtype selective”, and “selective ligand”. As the ligands were collected from different sources with their binding affinities measured under different assays and conditions, some level of variations in binding affinities is expected. Therefore, we tentatively selected compounds with binding affinity Ki < 1μM against a dopamine receptor as its ligands, and those with binding affinity Ki > 10μM as non-ligands. The 1μM to 10μM binding affinity gap between ligands and non-ligands was used for reducing

the possible influence of experimental binding affinity variations on the robustness of developed VS models. Some of the dopamine receptor ligands have been explicitly reported to be subtype selective or multi-subtype ligands, which can be used for testing the subtype selective capability of our developed VS models. Thus for subtypes with ≥20 subtype selective or ≥20 multi-subtype ligands, the corresponding ligands were used as independent testing datasets (a cut-off of 20 ligands was used to ensure the testing to be statistically meaningful).

We assembled 491 D1, 2202 D2, 1355 D3 and 1486 D4 ligands published before 2010 and 59 D1, 135 D2, 76 D3 and 29 D4 ligands published since 2010 with unspecified selectivity toward other subtypes, and 264 D1, 1577 D2, 631 D3 and 526 D4 non-ligands published before 2010 and 25 D1, 65 D2, 28 D3 and 33 D4 non-ligands published since 2010 with unspecified selectivity toward other subtypes. The collected pre-2010 ligands and non-ligands for each receptor subtype were used as positive and negative samples of the training dataset for developing VS models for that subtype. The collected non-ligands are insufficient to cover the vast non-ligand chemical space. Therefore, putative ligands for each receptor subtype were generated from the representative compounds of the compound families that contain no known ligand of that subtype by using the method described in our earlier studies [206]. A total of 65198 D1, 63687 D2, 62927 D3 and 63272 D4 putative non-ligands were generated and used in combination with known non-ligands as the negative samples of the training datasets. The collected post-2010 ligands and non-ligands were used as independent testing datasets for evaluating the performance of the developed VS models. These datasets are summarized in **Table 3-1**.

The use of pre-2010 and post-2010 compounds as training and testing datasets was intended to mimic the case of VS models being developed in 2010 and subsequently tested a few years later against newly discovered compounds. In view that such training and testing datasets and their developed models may not be easily reproduced and comparatively evaluated, we designed alternative training and testing datasets by randomly separating all ligands and non-ligands of a receptor subtype into approximately 10 compound-sets, with 9 compound-sets as a training dataset and the remaining 1 as a testing dataset (these training and testing datasets contain similar number of compounds as the corresponding ones developed from pre-2010 and post-2010 compounds). There are 10 sets of training and testing datasets for each subtype with each of the 10 compound-sets used as a testing dataset once, all of which were tested in this work. These alternative datasets are summarized in **Table 3-2**.

Table 3-2 Statistics of alternative training and testing datasets for D1, D2, D3 and D4 subtypes, and the performance of SVM models developed and tested by these datasets in predicting D1, D2, D3 and D4 ligands. SE, SP, Q and C are sensitivity, specificity, overall accuracy and Matthews correlation coefficient respectively.

Dopamine Receptor Subtype	Alternative dataset	Number of ligands/non-ligands in alternative training and testing dataset	VS performance on testing dataset			
			SE	SP	Q	C
D1	1	441/58914 and 50/6546	92.00%	99.89%	99.83%	0.79
	2	443/58914 and 48/6546	81.25%	99.93%	99.80%	0.73
	3	443/58914 and 48/6546	91.66%	99.93%	99.87%	0.84
	4	443/58914 and 48/6546	79.10%	99.95%	99.80%	0.73
	5	443/58914 and 48/6546	89.58%	99.89%	99.81%	0.77
	6	442/58914 and 49/6546	91.84%	99.98%	99.92%	0.9
	7	442/58914 and 49/6546	93.88%	99.90%	99.86%	0.83
	8	442/58914 and 49/6546	91.84%	99.98%	99.92%	0.9
	9	442/58914 and 49/6546	89.80%	99.98%	99.91%	0.88
	10	442/58914 and 49/6546	93.88%	99.92%	99.88%	0.84
	AVE		89.48%	99.94%	99.86%	0.82
	S.D		0.05127	0.00036	0.00048	0.06402
	S.E.M		0.01621	0.00011	0.00015	0.02025
D2	**Alternative dataset**	**Number of ligands/non-ligands in alternative training and testing dataset**	**SE**	**SP**	**Q**	**C**
	1	2178/58914 and 242/6546	88.84%	99.71%	99.32%	0.81
	2	2178/58914 and 242/6546	94.24%	99.79%	99.59%	0.88
	3	2178/58914 and 242/6546	93.00%	99.74%	99.50%	0.86
	4	2178/58914 and 242/6546	93.42%	99.77%	99.54%	0.87
	5	2178/58914 and 242/6546	91.82%	99.58%	99.33%	0.8
	6	2178/58914 and 242/6546	90.91%	99.74%	99.42%	0.84
	7	2178/58914 and 242/6546	94.21%	99.71%	99.51%	0.86
	8	2178/58914 and 242/6546	89.67%	99.71%	99.35%	0.82
	9	2178/58914 and 242/6546	89.67%	99.68%	99.32%	0.81
	10	2178/58914 and 242/6546	92.56%	99.74%	99.48%	0.86
	AVE		91.83%	99.72%	99.44%	84.10%
	S.D		0.01973	0.00058	0.00101	0.02885
	S.E.M		0.00624	0.00018	0.00032	0.00912
D3	**Alternative dataset**	**Number of ligands/non-ligands in alternative training and testing dataset**	**SE**	**SP**	**Q**	**C**
	1	1215/57564 and 135/6356	93.33%	99.78%	99.64%	0.83
	2	1215/57564 and 135/6356	92.59%	99.79%	99.65%	0.83
	3	1215/57564 and 135/6356	91.85%	99.79%	99.63%	0.83
	4	1215/57564 and 135/6356	91.11%	99.80%	99.63%	0.82
	5	1215/57564 and 135/6356	91.11%	99.85%	99.68%	0.84
	6	1215/57564 and 135/6356	93.33%	99.85%	99.72%	0.87
	7	1215/57564 and 135/6356	92.59%	99.83%	99.68%	0.85

	8	1215/57564 and 135/6356	90.37%	99.80%	99.60%	0.81
	9	1215/57564 and 135/6536	94.81%	99.86%	99.75%	0.88
	10	1215/57564 and 135/6536	94.81%	99.88%	99.78%	0.89
	AVE		92.59%	99.82%	99.68%	84.50%
	S.D		0.01521	0.00035	0.00058	0.02677
	S.E.M		0.00481	0.00011	0.00018	0.00847
D4	**Alternative dataset**	**Number of ligands/non-ligands in alternative training and testing dataset**	**SE**	**SP**	**Q**	**C**
	1	1332/57920 and 148/6380	93.91%	99.62%	99.49%	0.80
	2	1332/57920 and 148/6380	93.24%	99.69%	99.54%	0.81
	3	1332/57920 and 148/6380	91.89%	99.78%	99.60%	0.83
	4	1332/57920 and 148/6480	91.89%	99.74%	99.57%	0.82
	5	1332/57920 and 148/6480	91.89%	99.80%	99.63%	0.84
	6	1332/57920 and 148/6480	92.57%	99.89%	99.72%	0.87
	7	1332/57920 and 148/6480	91.21%	99.91%	99.71%	0.87
	8	1332/57920 and 148/6480	94.59%	99.87%	99.75%	0.89
	9	1332/57920 and 148/6380	89.86%	99.78%	99.55%	0.81
	10	1332/57920 and 148/6380	90%	99.78%	99.56%	0.81
	AVE		92.11%	99.79%	99.61%	83.50%
	S.D		0.01540	0.00090	0.00088	0.03136
	S.E.M		0.00487	0.00028	0.00028	0.00992

Dopamine receptor subtype selective ligands have been discovered and evaluated based on the criterion that each ligand binds to a specific subtype with at least ~10 fold higher binding affinity (Ki value) than that to another subtype [230]. Based on this criterion, we collected 97, 21, and 29 D1 selective ligands with > 10 fold higher binding affinity over D2, D3 and D4 respectively, 43, 37 and 63 D2 selective ligands over D1, D3 and D4 respectively, 48, 99 and 85 D3 selective ligands over D1, D2 and D4 respectively, and 27, 408 and 207 D4 selective ligands over D1, D2 and D3 respectively (**Table 3-3**). These subtype selective ligands were used as the positive samples to test subtype selectivity of our developed VS models.

Table 3-3 Datasets of our collected dopamine receptor D1, D2, D3 and D4 selective ligands against another subtype. The binding affinity ratio is the experimentally measured binding affinity to the second subtype divided by that to the first subtype: (Ki of the second subtype / Ki of the first subtype). This dataset was used as positive samples for testing subtype selectivity of our developed virtual screening models.

Dopamine receptor subtype	Selectivity against the second subtype	Number of subtype selective ligands against the second subtype	Range of binding affinity ratio	Mean of binding affinity ratio
D1	D2	97	10-4533	359
	D3	21	11-559	122
	D4	29	11-4600	770
D2	D1	43	10-3707	337
	D3	37	10-615	66
	D4	63	10-1851	113
D3	D1	48	17-38461	3863
	D2	99	10-6666	259
	D4	85	10-9111	950
D4	D1	27	13-4761	1315
	D2	408	10-10752	2962
	D3	207	10-51162	1175

The binding subtypes of a number of multi-subtype dopamine ligands have been explicitly reported [204, 205]. These ligands and their binding subtypes were selected based on the criterion that they bind to each subtype with binding affinity Ki < 1μM. We collected 4 groups of dual-subtype ligands (147 D1-D2, 4 D1-D3, 8 D1-D4, and 100 D3-D4 ligands), 2 groups of triple-subtype ligands (39 D1-D2-D3 and 2 D1-D2-D3 ligands), and 1 group of quadruple-subtype ligands (60 D1-D2-D3-D4 ligands). Four of these groups with > 10 ligands were selected as negative samples to test the ability of our developed VS models in predicting multi-subtype ligands (and thus the ability in separating subtype-selective and multi-subtype ligands) (**Table 3-4**). There are three other groups with high numbers of multi-subtype ligands (569 D2-D3, 276 D2-D4 and 402 D2-D3-D4 ligands). Separation of these groups of multi-subtype ligands from the training datasets would significantly compromise the structural diversity of the training

datasets. Therefore, these three groups were not removed from the training datasets. Inclusion of these groups in the training datasets does not enhance their subtype-selective signal. Instead they act as noise that tends to reduce the capability of the developed VS models in separating subtype-selective and multi-subtype ligands.

Table 3-4 Datasets of our collected dopamine receptor multi-subtype ligands. Four groups of this dataset were used as negative samples for testing subtype selectivity of our developed multi-label machine learning models.

Ligand Group	Binding Subtypes	Number of Ligands of Subtypes	Used as Testing Dataset
Dual Subtype Ligands	D1 and D2	147	Yes
	D1 and D3	4	No
	D1 and D4	8	No
	D3 and D4	100	Yes
Triple Subtype Ligands	D1, D2 and D3	39	Yes
	D1, D3 and D4	2	No
Quadruple Subtype Ligands	D1, D2, D3 and D4	60	Yes

ERα and ERβ ligands were collected in the same manner as that of dopamine receptor ligands using keyword combinations of “estrogen”, “estrogen receptor”, “ER”, “ER alpha”, “ER beta”, “ligand”, “binding”, “binder”, “subtype selective”, and “selective ligand”. We collected 1146 ERα and 1234 ERβ ligands (with unknown status about their subtype selectivity or multi-subtype binding) and 761 and 786 ERα and ERβ non-ligands, which together with 64013 and 60603 putative ERα and ERβ non-ligands (generated by the same procedure as the putative dopamine receptor subtype non-ligands) were used for training 2BR-SVM VS models using the same procedure as that of the alternative dataset version of dopamine receptor subtype selective VS models. There are 10 sets of randomly assembled training and testing datasets for each estrogen receptor

subtype with each of the 10 randomly generated compound-sets used as a testing dataset once, all of which were tested in this work. We also collected 40 and 55 ERα and ERβ selective ligands (with binding affinity ratios in the range of 10-2055 and 10-1143) and 63 ERα and ERβ multi-target ligands, which were used as independent testing datasets for testing the VS models. These datasets are summarized in **Table 3-5**.

Table 3-5 Statistics of the randomly assembled training and testing datasets for ERα and ERβ, and the performance of SVM models developed and tested by these datasets in predicting ERα and ERβ ligands. SE, SP, Q and C are sensitivity, specificity, overall accuracy and Matthews correlation coefficient respectively.

Estrogen Receptor Subtype	**Dataset**	**Number of ligands/non-ligands in alternative training and testing dataset**	**VS performance on testing dataset**			
			SE	**SP**	**Q**	**C**
ERα	1	1031/58296 and 115/6477	94.78%	99.92%	99.83%	0.90
	2	1031/58296 and 115/6477	96.52%	99.92%	99.86%	0.92
	3	1031/58296 and 115/6477	95.65%	99.90%	99.83%	0.90
	4	1031/58296 and 115/6477	93.04%	99.98%	99.86%	0.92
	5	1031/58296 and 115/6477	95.65%	99.83%	99.75%	0.87
	6	1031/58296 and 115/6477	94.78%	99.86%	99.77%	0.87
	7	1031/58296 and 114/6477	96.49%	99.85%	99.79%	0.88
	8	1032/58295 and 114/6478	95.61%	99.92%	99.85%	0.91
	9	1032/58295 and 114/6478	94.73%	99.92%	99.83%	0.90
	10	1032/58295 and 114/6478	93.86%	99.90%	99.82%	0.89
	AVE		95.11%	99.90%	99.82%	0.90
	S.D		0.01102	0.00043	0.00038	0.01838
	S.E.M		0.00349	0.00014	0.00012	0.00581
	Dataset	**Number of ligands/non-ligands in alternative training and testing dataset**	**SE**	**SP**	**Q**	**C**
ERβ	1	1110/54544 and 124/6060	92.94%	99.84%	99.67%	0.86
	2	1110/54544 and 124/6060	93.55%	99.88%	99.72%	0.89
	3	1110/54544 and 124/6060	93.55%	99.76%	99.61%	0.84
	4	1110/54544 and 124/6060	95.97%	99.80%	99.71%	0.88
	5	1111/54543 and 123/6061	94.31%	99.94%	99.81%	0.91
	6	1111/54543 and 123/6061	96.75%	99.88%	99.81%	0.92
	7	1111/54543 and 123/6061	96.75%	99.84%	99.77%	0.9
	8	1111/54543 and 123/6061	97.56%	99.74%	99.69%	0.88
	9	1111/54544 and 123/6060	94.31%	99.90%	99.77%	0.9
	10	1111/54544 and 123/6060	95.93%	99.88%	99.79%	0.91
	AVE		95.16%	99.85%	99.74%	0.89
	S.D		0.01620	0.00063	0.00066	0.02470
	S.E.M		0.00512	0.00020	0.00021	0.00781

3.2.2 Molecular representations

The 2D structures of our collected compounds were drawn by using Chemdraw or from the ChEMBLdb [125] and Pubchem [210] databases. Each compound was represented by 98 molecular descriptors (**Table 3-6**) computed by using own developed MODEL program [231]. These 98 descriptors have been selected in our previous studies for developing VS models of a variety of target classes including GPCR ligands to screen large chemical libraries such as Pubchem compounds [99, 206, 232]. Although the structures of the binders of one target or subtype can be very different from those of another target or subtype, each binders set plus the representatives of the non-binders cover the same chemical space defined by the 13.56 million Pubchem compounds. Therefore, the same set of molecular descriptors was used in this work as well as our previous works [99, 232].

Table 3-6 List of 98 molecular descriptors computed by using our own developed MODEL program.

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

3.2.3 Support vector machines

SVM is based on the structural risk minimization principle for minimizing both training and generalization error [164]. In linearly separable cases, SVM constructs a hyper-plane to separate active and inactive classes of compounds with

a maximum margin. In nonlinearly separable cases, which frequently occur in classifying compounds of diverse structures [99, 206, 232], SVM maps the input vectors into a higher dimensional feature space by using the Radial Basis Function (RBF) kernel function. This kernel function has been extensively used and consistently shown better performance than other kernel functions [176-178]. In the high dimensional space, linear SVM can be applied for classifying the active and inactive compounds. For the parameters, a hard margin C= 100000 was used and σ= 0.4-0.6 were determined from 5 fold cross validation studies.

3.2.4 Combinatorial SVM method

In combinatorial strategy, SVM models for each receptor subtype are separately constructed, which are subsequently used for parallel screening against each individual subtype to find compounds that only bind to one of the subtypes (putative subtype selective ligands) or simultaneously bind to multiple subtypes (putative subtype non-selective ligands) [99, 232].

3.2.5 Two-step Binary relevance SVM method

Subtype selective ligands were selected by two steps. In the first step, a high performance SVM model was developed for each receptor subtype to select ligands of that subtype regardless of their selectivity towards other subtypes. The high performance in selecting ligands of a subtype was achieved by using comprehensive sets of known ligands and putative non-ligands of the

corresponding receptor to train the respective SVM model [206]. In the second step, the Binary relevance (BR) method [215] was used for more refined selection of subtype selective ligands from the putative ligands selected in the first step. BR is a popular multiple binary classification method that transforms the original N-label dataset into N pairs of datasets with samples of each label as positive dataset and samples of the other N-1 labels as negative dataset [215].

3.2.6 Multi-label K nearest neighbor method

ML-kNN implemented in the Mulan software package [215] was used in this work. ML-kNN [216] extends traditional kNN method to solve the multi-label problem. In the first step, ML-kNN classifies a compound $\boldsymbol{x}$ by linking it to the known ligand or non-ligand $\boldsymbol{x}_i$ in the training dataset with closest Euclidean distance [180]. In the second step, statistical information, i.e. prior and posterior probabilities for the frequency of each label within the k nearest neighbors, is gained from the label sets of these neighboring ligands. In the third step, maximum a posteriori (MAP) principle is used to determine the label set for the unknown ligands. The default parameters in Mulan package were used in this work.

3.2.7 The random k-labelsets decision tree method

RAkEL-DT implemented in the Mulan software package [215] was used in this work. The random k-labelsets (RAkEL) method [217] constructs an ensemble of

label powerset (LP) classifiers. LP is a transformation method which considers each unique set of labels existed in multi-label training set as new single label. Since RAkEL is a transformation-based algorithm and it accepts a single-label classifier as a parameter, decision tree C4.5 algorithm was used for this purpose. C4.5 decision tree is a branch-test-based classifier [233]. A branch in a decision tree corresponds to a group of classes and a leaf represents a specific class. A decision node specifies a test to be conducted on a single attribute value, with one branch and its subsequent classes as possible outcomes of the test. C4.5 decision tree uses recursive partitioning to examine every attribute of the data and to subsequently rank them according to their ability to partition the remaining data, thereby constructing a decision tree. The default parameters in Mulan package were used in this work.

3.2.8 Virtual screening model development, parameter determination and performance evaluation

All VS models for each dopamine receptor subtype were trained from the training datasets in **Table 3-1**. The parameters were determined by 5-fold cross validation (CV) tests, and the performance of these VS models was evaluated by using the independent testing datasets in **Table 3-1**. In each 5-fold CV test, the training dataset was divided into 5 groups of approximately equal number of positive samples and equal number of negative samples, with 4 groups used for training and 1 group used for testing the model. There are five such sets each with one unique group used as a testing set, from which five prediction models can be

constructed. VS models were developed at different parameters. The parameters with the best overall 5-fold CV performance were selected for developing the final VS models.

In 5-fold cross validation studies, the inhibitor and non-inhibitor prediction accuracies are given by sensitivity and specificity respectively. Prediction accuracies have also been frequently measured by overall prediction accuracy (*Q*) and Matthews correlation coefficient (*C*). In the large database screening tests, the yield and false-hit rate are given respectively. The detailed performance evaluation is described in Chapter 2 (section 2.4).

3.2.9 Determination of similarity level of a compound against dopamine receptor ligands in a dataset

The similarity level of a compound *i* with respect to the ligands of a dataset can be determined by using the Tanimoto coefficient *sim(i,j):* [235].

$$sim(i,j) = \frac{\sum_{d=1}^{l} x_{di} x_{dj}}{\sum_{d=1}^{l} (x_{di})^2 + \sum_{d=1}^{l} (x_{dj})^2 - \sum_{d=1}^{l} x_{di} x_{dj}} \qquad (3)$$

where x_{di} represents a molecular fingerprint of compound *i* (there are 882 fingerprints calculated from the PaDEL-Descriptors program [141], *l* is the number of molecular fingerprints, *j* is the index of the ligand in the dataset most similar to compound *i*. The compound *i*. is assigned into one of the ten similarity levels based on its computed *sim(i,j)* values: 0.9-1, 0.8-0.9, 0.7-0.8, 0.6-0.7,

0.5-0.6, 0.4-0.5, 0.3-0.4, 0.2-0.3, 0.1-0.2, and 0-0.1. Compounds are typically considered to be highly similar if *sim(i,j)* is no less than 0.8 or 0.9 [183, 184].

3.2.10 Determination of dopamine receptor subtype selective features by feature selection method

Molecular features important for dopamine receptor subtype selective ligands were probed by using a feature selection method, recursive feature elimination (RFE) method, extensively used in selecting molecular features of compounds of specific pharmacodynamic and pharmacokinetic properties [138]. In this approach, the level of contribution of individual molecular descriptor to SVM classification of ligands of a subtype against ligands of other subtypes was ranked and the top-ranked ones were selected based on the evaluation of the variation of the SVM objective function *J* caused by the removal of an individual descriptor [236]. The variation $DJ(i)$ due to the removal of a descriptor i is computed by $DJ(i) = \frac{1}{2}\frac{\partial^2 J}{\partial w_i^2}(Dw_i)^2$ with the weight variation determined by $Dw_i = w_i$. In this work, Gaussian kernels were used for developing SVM models. In this case, $DJ(i) = (1/2)\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{H}\boldsymbol{\alpha} - (1/2)\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{H}(-i)\boldsymbol{\alpha}$, where H is the matrix with elements $y_i y_j \exp(-\|x_i - x_j\|^2/(2\sigma^2))$, $H(-i)$ is the matrix computed by the same method as matrix H but with its i-th component removed, y_i is the vector composed of molecular descriptors, 1 is an m dimensional identity vector (m is the number of compounds in a training dataset), and the component of vector α is kept in the range of $0 \leq \alpha_k \leq C$.

The computational procedure for selecting subtype selective features is as follows: For a specific subtype, the corresponding SVM model developed in the second step of the 2SBR-SVM method is processed by iteratively evaluating and eliminating molecular descriptors at different parameter σ values based on 5-fold cross-validation. In the first step, for a fixed σ, the SVM is trained by using the complete set of descriptors (feature set). The second step is to compute the ranking criterion score $DJ(i)$ for every existing descriptor. All the computed $DJ(i)$ is then ranked in descending order. The third step is to remove the m descriptors with smallest criterion scores (m=4 in this work). In the fourth step, the SVM is retrained by using the remaining molecular descriptors and a new prediction accuracy of 5-fold cross-validation is computed. The second to fourth steps are repeated for multiple-iterations until all descriptors are removed. For another fixed σ, the first to fourth steps are repeated.

3.3 Results and discussion

3.3.1 5-fold cross-validation tests

The results of 5-fold CV tests of the SVM models of D1, D2, D3 and D4 ligands are shown in **Table 3-7**. Overall, the sensitivity, specificity, overall accuracy and the Matthews correlation coefficients of the best performing SVM models are in the range of 87.8%-95.3%, 99.6%-99.9%, 99.3%-99.8%, and 0.74-0.90 respectively. These results are comparable to those of our earlier studies [232], indicating that the SVM models for dopamine receptor subtypes have similar prediction capability as those for other target classes. The VS models with the best 5-fold CV performance were further tested on independent sets of

dopamine receptor ligands and non-ligands published since 2010 and not in the training datasets, which are also shown in **Table 3-7**. The sensitivity, specificity and overall accuracy are in the range of 71.2%-89.7%, 61.5%-76.0% and 71.4%-82.7% respectively. The sensitivity is substantially smaller than that of 5-fold CV tests. This is because many of the post-2010 ligands in the independent datasets are structurally different from those of the pre-2010 ligands in the training datasets. As shown in **Table 3-1**, 16.3%-34.5% of the post-2010 ligands are outside the chemical families of pre-2010 ligands in the training datasets. The specificity is also significantly smaller than that of the 5-fold CV tests. This is partly because many non-ligands have weak (Ki 10-50μM) binding activity and may thus be difficult to be separated from the ligands.

Table 3-7 Results of 5-fold cross validation (CV) tests of SVM models in predicting D1, D2, D3 and D4 ligands. SE, SP, Q and C are sensitivity, specificity, overall accuracy and Matthews correlation coefficient respectively.

Dopamine Receptor Subtype	**5-fold CV Tests for Parameter Selection Based on the Training Datasets in Table 1**						**5-fold CV Tests for Performance Evaluation Based on the Independent Testing Datasets in Table 1**			
D1	**Fold**	**Number of ligands/non-ligands**	**SE**	**SP**	**Q**	**C**	**Number of ligands/ non-ligands**	**SE**	**SP**	**Q**
	1	99/13092	91.92%	99.87%	99.81%	0.77	59/25	71.19%	76.00%	72.62%
	2	99/13093	88.78%	99.91%	99.83%	0.78	59/25	72.88%	72.00%	71.43%
	3	98/13093	88.78%	99.87%	99.79%	0.74	59/25	71.19%	72.00%	71.43%
	4	98/13092	87.76%	99.88%	99.79%	0.74	59/25	71.19%	72.00%	71.43%
	5	97/13093	87.76%	99.92%	99.83%	0.78	59/25	71.19%	72.00%	71.43%
	AVE		89.00%	99.89%	99.81%	0.76		71.53%	72.80%	71.67%
	S.D		0.01710	0.00024	0.0002	0.02049		0.00756	0.01789	0.00532
	S.E.M		0.00765	0.00011	0.00009	0.00917		0.00338	0.008	0.00238
D2	**Fold**	**Number of ligands/non-ligands**	**SE**	**SP**	**Q**	**C**	**Number of ligands/ non-ligands**	**SE**	**SP**	**Q**
	1	441/13092	92.74%	99.66%	99.44%	0.83	135/65	86.67%	61.54%	78.50%
	2	441/13092	94.10%	99.70%	99.52%	0.86	135/65	88.15%	63.08%	80.00%
	3	440/13093	93.12%	99.68%	99.47%	0.84	135/65	86.67%	63.08%	79.00%
	4	440/13093	91.82%	99.67%	99.42%	0.82	135/65	85.93%	67.69%	80.00%
	5	440/13092	91.82%	99.58%	99.33%	0.80	135/65	85.93%	61.54%	78.00%
	AVE		92.72%	99.66%	99.44%	0.83		86.67%	63.39%	79.10%
	S.D		0.00960	0.00046	0.00070	0.02236		0.00906	0.02526	0.00894

	S.E.M		0.00429	0.00021	0.00031	0.01		0.00405	0.01130	0.004
D3	**Fold**	**Number of ligands/non-ligands**	**SE**	**SP**	**Q**	**C**	**Number of ligands/ non-ligands**	**SE**	**SP**	**Q**
	1	271/12712	90.77%	99.75%	99.56%	0.80	76/28	86.84%	64.29%	79.81%
	2	271/12712	93.73%	99.79%	99.66%	0.84	76/28	89.47%	67.86%	82.69%
	3	271/12712	92.99%	99.77%	99.63%	0.83	76/28	86.84%	64.29%	79.81%
	4	271/12711	89.67%	99.82%	99.61%	0.82	76/28	84.21%	67.86%	79.81%
	5	271/12711	95.20%	99.88%	99.78%	0.90	76/28	89.47%	64.29%	81.73%
	AVE		92.47%	99.80%	99.65%	0.84		87.37%	65.72%	80.77%
	S.D		0.02238	0.00051	0.00082	0.03768		0.022	0.01955	0.0136
	S.E.M		0.01001	0.00023	0.00037	0.01685		0.00984	0.00875	0.00608
D4	**Fold**	**Number of ligands/non-ligands**	**SE**	**SP**	**Q**	**C**	**Number of ligands/ non-ligands**	**SE**	**SP**	**Q**
	1	297/12760	95.29%	99.71%	99.61%	0.84	29/33	89.66%	72.73%	80.65%
	2	297/12760	93.27%	99.76%	99.61%	0.84	29/33	86.21%	63.64%	74.19%
	3	297/12760	94.28%	99.80%	99.67%	0.86	29/33	89.66%	63.64%	75.81%
	4	297/12759	93.94%	99.79%	99.66%	0.85	29/33	86.21%	63.64%	74.19%
	5	297/12759	94.61%	99.76%	99.65%	0.85	29/33	86.21%	63.64%	74.19%
	AVE		94.28%	99.76%	99.64%	0.848		87.59%	65.46%	75.81%
	S.D		0.00752	0.00035	0.00028	0.00837		0.0189	0.04065	0.02794
	S.E.M		0.00337	0.00015	0.00013	0.00374		0.00845	0.01818	0.01249

The VS performance of the SVM VS models developed by the 10 sets of alternative training and testing datasets is provided in **Table 3-2**. The sensitivity, specificity, overall accuracy and the Matthews correlation coefficients of these SVM models in classifying dopamine receptor subtype ligands and non-ligands are in the range of 79.1%-94.8%, 99.6%-99.9%, 99.3%-99.9%, and 0.73-0.90 respectively, which are very similar to those of the SVM models developed by pre-2010 and tested by post-2010 compounds. A further analysis of structures of the randomly assembled datasets and those of the chronologically assembled datasets showed that most of the active and inactive scaffolds are mutually represented on both sides because of the significant structural diversity in these datasets. Therefore, the VS performance of SVM models developed by chronologically assembled datasets can be compared with those models developed by using datasets assembled by conventional approach.

3.3.2 Applicability domains of the developed SVM VS models

Our SVM VS models for each dopamine receptor subtype were developed by using known ligands and non-ligands of the subtype, and the putative non-ligands composed of representative compounds of all of the compound families in the Pubchem chemical space that contain no known ligand of the subtype. Theoretically, these VS models are expected to be applicable in the chemical space defined by the known ligands, known non-ligands, and the 13.56M Pubchem compounds. If this is true, in addition to good predictive performance on the known ligands, these VS models are expected to consistently identify very

small percentages of Pubchem compounds as subtype selective ligands regardless of their similarity levels to the known ligands. Alternatively, if the applicability domain of these models covers limited chemical space around known ligands, then the number of identified Pubchem compounds may increase substantially beyond the applicability domain (i.e. at lower similarity levels). To determine the applicability domain of each SVM VS model, we divided 13.56M PubChem compounds into groups of 10 similarity levels with respect the known ligands of each receptor subtype (defined in the methods section), and then monitored if the number of the SVM identified PubChem compounds significantly increases at higher similarity levels. As shown in **Table 3-8**, the percentages of identified Pubchem compounds for all four receptor subtypes (0.0489%-0.0521% for D1, 0.131%-0.135% for D2, 0.143%-0.147% for D3, and 0.157%-0.160% for D4 respectively) are consistently small and show little variations at different similarity levels. This suggests that the applicability domains of our SVM VS models likely cover the chemical space defined by the known ligands, known non-ligands and the PubChem compounds.

Table 3-8 Numbers of Pubchem compounds at different similarity levels with respect to known ligands of each dopamine receptor subtype, and percent of these compounds identified by SVM VS model as subtype selective ligands.

Dopamine receptor subtype		Similarity level with respect to known ligands of the subtype defined by Tanimoto similarity score									
		0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1
D1	Number of Pubchem compounds at the similarity level	366852	1238930	2210766	3832638	3652430	781974	384551	355589	339499	389378
	Percent of these Pubchem compounds identified as subtype selective ligand	0.0499%	0.0489%	0.0498%	0.0521%	0.0509%	0.0493%	0.0486%	0.0515%	0.0510%	0.0507%
D2	Number of Pubchem compounds at the similarity level	477873	1111819	1464190	1707149	3026529	2593708	892995	659690	812545	806109
	Percent of these Pubchem compounds identified as subtype selective ligand	0.1306%	0.1320%	0.1322%	0.1350%	0.1311%	0.1306%	0.1303%	0.1326%	0.1309%	0.1310%

D3	Number of Pubchem compounds at the similarity level	770711	1497979	2325005	3232481	1718412	896213	664517	662908	812545	650036
	Percent of these Pubchem compounds identified as subtype selective ligand	0.1445%	0.1471%	0.1434%	0.1475%	0.1467%	0.1456%	0.1477%	0.1469%	0.1470%	0.1473%
D4	Number of Pubchem compounds at the similarity level	947701	1348342	2672549	2548656	2350749	942874	778756	662908	733704	566368
	Percent of these Pubchem compounds identified as subtype selective ligand	0.1601%	0.1593%	0.1579%	0.1568%	0.1580%	0.1591%	0.1588%	0.1576%	0.1582%	0.1579%

3.3.3 Prediction performance on dopamine receptor subtype selective and multi-subtype ligands

The performance of our new method 2SBR-SVM and that of the three previously used methods Combi-SVM, ML-kNN and RAkEL-DT in predicting dopamine subtype selective ligands was determined as follows: For each set of dopamine receptor subtype selective ligands against another subtype, the developed VS model of the subtype and that of the second subtype were both used to screen these ligands. The percentage of these ligands selected by the first model but not by the second model was used to measure the performance of the VS models in selecting subtype selective ligands. The relevant results are shown in **Table 3-9**.

Table 3-9 The performance of our new method 2SBR-SVM and that of previously used methods Combi-SVM, ML-kNN and RAkEL-DT in predicting dopamine receptor subtype selective ligands.

Dopamine receptor subtype	Selectivity against the second subtype	Number of subtype selective ligands	Percent of subtype selective ligands predicted as subtype selective with respect to the second subtype			
			Combi-SVM	ML-kNN	RAkEL-DT	2SBR-SVM
D1	D2	97	13.40%	30.93%	75.26%	86.60%
	D3	21	23.81%	23.81%	47.62%	66.67%
	D4	29	17.24%	58.62%	44.83%	65.52%
	Average		18.15%	37.79%	55.90%	72.93%
D2	D1	43	55.81%	62.79%	69.77%	93.02%
	D3	37	16.22%	21.62%	62.16%	81.08%
	D4	63	14.29%	39.68%	30.16%	82.54%
	Average		28.77%	41.36%	54.03%	85.55%
D3	D1	48	72.92%	87.50%	85.42%	56.25%
	D2	99	22.22%	26.26%	50.51%	51.52%
	D4	85	17.65%	31.76%	22.35%	50.59%
	Average		37.60%	48.51%	52.76%	52.79%
D4	D1	27	74.07%	70.37%	85.19%	82.50%
	D2	408	33.33%	28.43%	57.60%	88.00%
	D3	209	26.79%	24.40%	45.46%	83.73%
	Average		44.73%	41.07%	62.75%	84.74%

As shown in **Table 3-9**, the three previously used methods showed mostly moderate and in minority cases good performance in predicting dopamine receptor subtype selective ligands. Specifically, 13.4%-23.8%, 14.3%-55.8%, 17.7%-77.9% and 26.8%-74.1% of the D1, D2, D3 and D4 selective ligands were correctly predicted by Combi-SVM as subtype selective ones. ML-kNN showed better performance, correctly predicting 23.8%-58.6%, 21.6%-62.8%, 26.3%-87.5% and 24.4%-70.4% of the D1, D2, D3 and D4 selective ligands as subtype selective ones. The RAkEL-DT method achieved the best performance among the three methods, correctly predicting 44.8%-75.3%, 30.2%-69.8%, 22.4%-85.4% and 45.5%-85.2% of the D1, D2, D3 and D4 selective ligands as subtype selective ones. On the other hand, our new method 2BR-SVM produced significantly improved performance, correctly predicting 66.5%-86.6%, 81.1%-93.0%, 50.6%-56.3% and 82.5%-88.0% of the D1, D2, D3 and D4 selective ligands as subtype selective ones. This suggests that our two-step strategy with one step focusing on subtype binding and another on selectivity works more effectively than the three previously used methods in predicting dopamine receptor subtype selective ligands.

The improved subtype selective performance of the 2BR-SVM method arises from its more rigorous evaluation of minor structural and physicochemical differences of subtype selective ligands. Comparative structural analysis has shown that some D2 selective and D3 selective ligands are highly similar in structure and interact with their respective subtypes in a very similar binding mode with some functional group adopting different orientation at sites of non-conserved residues [230]. Such minor differences may not be adequately

distinguished by conventional VS models developed by training datasets with inadequate representation of ligands of other subtypes, but may be distinguished by 2BR-SVM method with additional models developed by training datasets with sufficient representation of other subtypes.

The performance in predicting dopamine subtype selective ligands is measured not only by the capability in selecting subtype selective ligands, but also on the ability in differentiating them from multi-subtype ligands. Good prediction on subtype selective ligands needs to be complemented by equally good performance in predicting multi-subtype ligands as subtype non-selective ones. This performance was determined as follows: For each set of multi-subtype ligands (e.g. triple-subtype D1, D2 and D3 ligands), the VS models of all of the corresponding subtypes (e.g. D1, D2 and D3) were used to screen the multi-subtype ligands in the set. The percentage of these ligands selected by the model of more than one subtype was used to measure the performance of the VS models in predicting multi-subtype ligands as subtype non-selective ligands. The results are shown in **Table 3-10**.

Table 3-10 The performance of our new method 2SBR-SVM and that of previously used methods Combi-SVM, ML-kNN and RAkEL-DT in predicting dopamine receptor multi-subtype ligands as non-selective ligands.

Ligand Group	Binding subtypes	Number of Multi-Subtype Ligands	Percent of multi-subtype ligands predicted as non-selective ligands			
			Combi-SVM	ML-kNN	RAkEL-DT	2SBR-SVM
Dual Subtype Ligands	D1 and D2	147	68.02%	31.97%	35.37%	76.19%
	D3 and D4	100	83.0%	37.0%	39.0%	81.0%
Triple Subtype Ligands	D1, D2 and D3	39	76.92%	28.2%	33.33%	71.79%
Quadruple Subtype Ligands	D1, D2, D3 and D4	60	75.42%	36.67%	38.75%	71.67%

Of the three previously used methods, Combi-SVM showed the best performance in predicting dopamine receptor multi-subtype ligands as subtype non-selective ones, correctly predicting 68.0%, 83.0%, 76.9% and 75.4% of the D1-D2, D3-D4, D1-D2-D3 and D1-D2-D3-D4 multi-subtype ligands as subtype non-selective ones. On the other hand, only 32.0%, 37.0%, 28.2% and 36.7% of the D1-D2, D3-D4, D1-D2-D3 and D1-D2-D3-D4 multi-subtype ligands were predicted by ML-kNN as subtype non-selective ones, and only 35.4%, 39.0%, 33.3% and 38.8% of the D1-D2, D3-D4, D1-D2-D3 and D1-D2-D3-D4 multi-subtype ligands were predicted by RAkEL-DT as subtype non-selective ones. Hence, the better performance of ML-kNN and RAkEL-DT over Combi-SVM in predicting subtype selective ligands is off-set by the poorer performance in predicting multi-subtype ligands as subtype non-selective. Taken these two indicators together, Combi-SVM appears to show better overall performance in predicting subtype selective and subtype non-selective ligands than the ML-kNN and RakEL-DT methods.

The performance of our new method 2SBR-SVM in predicting dopamine receptor subtype non-selective ligands is similar to that of Combi-SVM, correctly predicting 76.2%, 81.0%, 71.8% and 71.7% of the D1-D2, D3-D4, D1-D2-D3 and D1-D2-D3-D4 multi-subtype ligands as subtype non-selective ones. Thus, our new method maintains the same performance level as that of the best performing method of the previously used methods in predicting dopamine receptor subtype non-selective ligands. The lack of improvement by our new method in predicting dopamine receptor subtype non-selective ligands may be partly due to the quality of training datasets. It is noted that three groups of multi-subtype ligands were included as positive samples in the training datasets, which likely affect the ability of the SVM models in predicting multi-subtype ligands as subtype non-selective ones.

3.3.4 Virtual screening performance in searching large chemical libraries

The virtual screening performance of our new method 2SBR-SVM and our previously developed method Combi-SVM was evaluated by using them to screen 13.56M Pubchem compounds, 168,016 MDDR compounds and 657,736 ChEMBLdb compounds to determine the numbers of Pubchem, MDDR, and ChEMBLdb compounds predicted as D1, D2, D3 and D4 selective ligands, which are shown in **Table 3-11**. For comparison, **Table 3-11** also includes the results of

SVM (single label) in identifying Pubchem compounds as putative D1, D2, D3 and D4 ligands regardless of their possible binding with another subtype. In screening Pubchem compounds, the number of D1, D2, D3 and D4 selective virtual hits identified by 2SBR-SVM and the corresponding virtual hit rate is 650 and 0.0048%, 1132 and 0.0083%, 1498 and 0.011%, and 1961 and 0.015% respectively, which is significantly smaller than those identified by Combi-SVM. The number of D1, D2, D3 and D4 selective virtual hits identified by Combi-SVM and the corresponding virtual hit rate is 4948 and 0.037%, 10080 and 0.074%, 6055 and 0.045%, and 9180 and 0.068% respectively. The number of virtual hits identified by Combi-SVM is nonetheless substantially smaller than that of single label SVM. The number of D1, D2, D3 and D4 selective virtual hits identified by single label SVM and the corresponding virtual hit rate is 6798 and 0.05%, 17786 and 0.13%, 19813 and 0.15%, and 21444 and 0.16% respectively. Some of the identified virtual hits are possible subtype selective ligands. Therefore the true false hit rates of the tested VS models are likely smaller than the computed virtual hit rates. The false hit rates of 2SBR-SVM in screening 13.56 million Pubchem compounds can then be estimated as $\leq$0.0048%, $\leq$0.0083%, $\leq$0.011% and $\leq$0.015% for D1, D2, D3 and D4 selective ligands respectively. Therefore, 2SBR-SVM produced very low false hit rates in screening large chemical libraries as well as good performance in selecting subtype selective ligands.

Table 3-11 Virtual screening performance of our new method 2SBR-SVM and that of our previously used method Combi-SVM in scanning 168,016 MDDR compounds and 657,736 ChEMBLdb compounds, and 13.56 million Pubchem compounds. For comparison, the results of single label SVM, which identify putative subtype binding ligands regardless of their possible binding to another subtype, are also included.

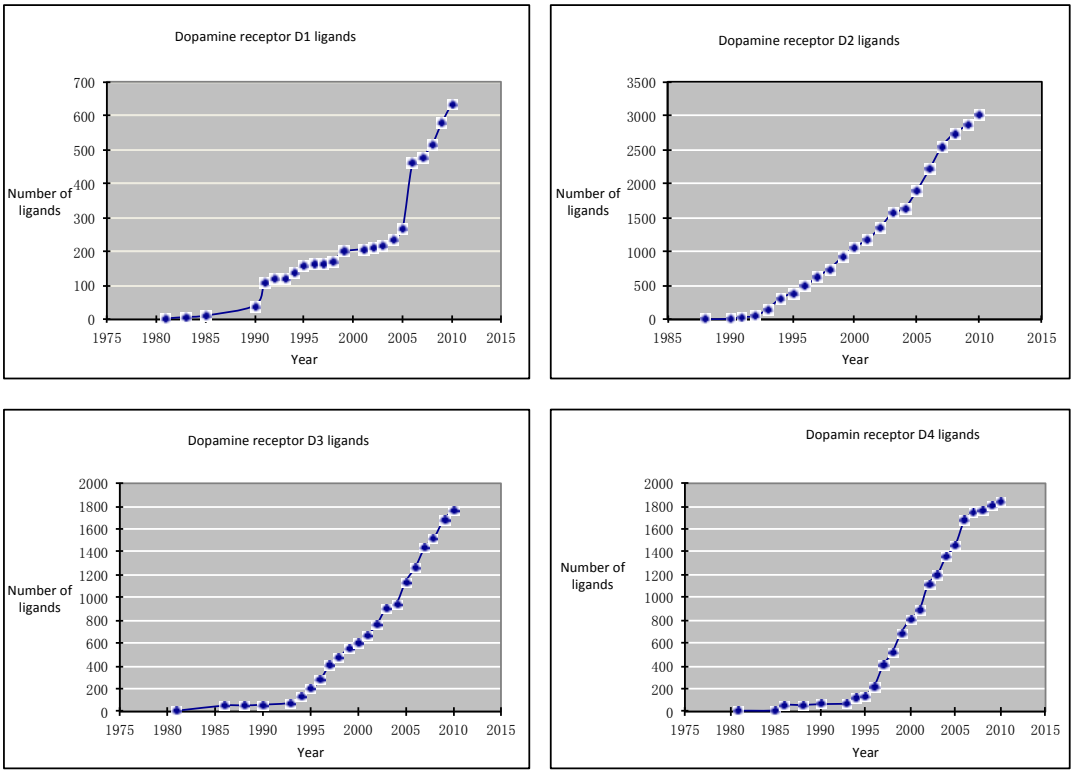
Dopamine receptor subtype	Method	Number and Percent of the 13.56M PubChem Compounds Identified as subtype selective ligands	Number and Percent of the 168,016 MDDR Compounds Identified as subtype selective ligands	Number and Percent of the 657,736 ChemBLdb Compounds Identified as subtype selective ligands
D1	SVM (Single Label)	6798(0.0501%)	463(0.28%)	1034(0.16%)
	Combi-SVM	4948(0.0365%)	383(0.23%)	755(0.11%)
	2SBR-SVM	650(0.0048%)	140(0.08%)	355(0.05%)
D2	SVM (Single Label)	17786(0.1312%)	1105(0.66%)	3208(0.49%)
	Combi-SVM	10080(0.0743%)	712(0.42%)	2023(0.31%)
	2SBR-SVM	1132(0.0083%)	108(0.06%)	686(0.10%)
D3	SVM (Single Label)	19813(0.1461%)	1149(0.68%)	3057(0.46%)
	Combi-SVM	6055(0.0447%)	679(0.40%)	1894(0.29%)
	2SBR-SVM	1498(0.0110%)	156(0.09%)	687(0.10%)
D4	SVM (Single Label)	21444(0.1581%)	1160(0.69%)	3489(0.53%)
	Combi-SVM	9186(0.0677%)	790(0.47%)	2579(0.39%)
	2SBR-SVM	1961(0.0145%)	134(0.08%)	907(0.14%)

As shown in **Table 3-11**, in screening MDDR and ChEMBLdb compounds, 2SBR-SVM as well as Combi-SVM and single label SVM produced reasonably low virtual hit rates that are in the range of 0.06%-0.09% and 0.05%-0.14% respectively, which are 10 fold higher than those in screening Pubchem compounds. MDDR and ChEMBLdb compounds as a collection of bioactive agents tend to be structurally closer to the dopamine receptor ligands than many Pubchem compounds that consist of high percentage of inactive compounds. Therefore, it tends to be more difficult for 2SBR-SVM to distinguish dopamine receptor ligands from some of the non-ligands in MDDR and ChEMBLdb databases, leading to higher virtual-hit rates. The virtual hit rates of 2SBR-SVM in screening MDDR and ChEMBLdb compounds are substantially (2-10 fold)

smaller than those of Combi-SVM and single label SVM, which suggests that 2SBR-SVM is capable of achieving lower false-hit rate in screening bioactive compounds than more conventional SVM methods.

Although it is unclear how many true D1, D2, D3 and D4 selective ligands are contained in Pubchem database. Some crude estimates can be made. As shown in **Table 3-1** and **Table 3-3**, the number of known ligands of a dopamine receptor subtype is in the range of 550-2337, and the number of known dopamine receptor subtype selective ligands is in the range of 21-408. The known subtype selective ligands are approximately 10 fold less in numbers than the known ligands of a subtype. While the numbers of the published D1, D2, D3, and D4 ligands continuously increase through the years (**Figure 3-1**), there are signs of significant reduction of the growth rates at the level of 2000-3000 ligands. These trends tend to project the existence of no more than several thousand undiscovered ligands for each dopamine receptor subtype in the chemical space defined by the Pubchem, MDDR and ChEMBLdb compounds. Hence, the number of subtype selective virtual hits identified by 2SRB-SVM is closer to the estimated upper limit of undiscovered dopamine receptor subtype ligands than those of Combi-SVM and single label SVM.

Figure 3-1 Number of published dopamine receptors D1, D2, D3 and D4 ligands from 1975 to present.



3.3.5 Dopamine receptor subtype selective features

The molecular descriptors important for distinguishing the ligands of every dopamine receptor subtype and the ligands of other subtypes were determined by using the feature selection method [138] outlined in the method section, which are provided in **Table 3-12**. The top-ranked D1 selective descriptors are number of O atoms, sum of Estate of atom type dssC, ssO and ssNH, graph-theoretical shape coefficient, and sum of H Estate of atom type HsNH2. These descriptors are consistent with the D1 selective features derived from a pharmacophoric model that includes positive nitrogens (linked to ssNH, HsNH2), hydrogen bond acceptor

(linked to O, ssO) and donor (linked to ssNH, HsNH2) [202]. The top-ranked D2 selective descriptors are number of H-bond acceptor, sum of H Estate of atom types HaaNH and HCsats, and sum of Estate of atom type dssC, aasC and aaNH. These are consistent with a CoMSIA based analysis that suggests that D2 selectivity is determined by hydrogen bond acceptor (linked to H-bond acceptor) and donor (linked to HaaNH), hydrophobic (linked to HCsats, dssC, aasC), and electrostatic (linked to HaaNH, aaNH) factors [198]. These are also consistent with the conclusion from a pharmacophoric model that two hydrogen acceptors or one hydrogen acceptor plus one donor are critically important for D2 selectivity of some ligands [202].

Table 3-12 Top-ranked molecular descriptors for distinguishing dopamine receptor subtype D1, D2, D3 or D4 selective ligands selected by RFE feature selection method.

Dopamine receptor subtype	**Top-ranked molecular descriptors for distinguishing subtype selective ligands and ligands of other subtypes**
D1	Number of O atoms, Sum of Estate of atom type dssC, Sum of Estate of atom type ssO, Sum of Estate of atom type ssNH, Graph-theoretical shape coefficient, Sum of H Estate of atom type HsNH2
D2	Number of H-bond acceptor, Sum of H Estate of atom type HaaNH, Sum of H Estate of atom type HCsats, Sum of Estate of atom type dssC, Sum of Estate of atom type aasC, Sum of Estate of atom type aaNH
D3	Sum of Estate of atom type dsCH, Sum of H Estate of atom type HsOH, Sum of H Estate of atom type HCsats, Sum of Estate of atom type aaaC, Sum of Estate of atom type sOH, Number of H-bond donnor
D4	Molecular path count of length 2, Sum of Estate of atom type ssCH2, 3th order Kier shape index, Topological radius, Sum of Estate of atom type aasC, Kier Molecular Flexibility Index

The top-ranked D3 selective descriptors are sum of Estate of atom type dsCH, aaaC and sOH, sum of H Estate of atom type HsOH and HCsats, and number of H-bond donor. These are consistent with the conclusions from several CoMSIA models that correlate D3 selectivity with specific hydrogen bond donor (linked to H-bond donor, sOH, HsOH), hydrophobic (linked to dsCH, aaaC), and electrostatic (linked to sOH, HsOH) factors [198, 203]. Moreover, a study of a D3 selective ligand further shows that hydrogen bonding from a hydroxyl group is important for conferring D3 selectivity [198]. The top-ranked D4 selective descriptors are molecular path count of length 2, sum of Estate of atom type ssCH2 and aasC, 3th order Kier shape index, topological radius, and Kier molecular flexibility index. These are consistent with a report that D4 selectivity is strongly influenced by the geometry and orientation of specific chemical groups (linked to molecular path count of length 2, 3th order Kier shape index, topological radius, and Kier molecular flexibility index) [197]. The consistency of our selected molecular descriptors and the literature-reported features for D1, D2, D3, and D4 selectivity suggests that the subtype selective molecular descriptors selected by our feature selection method may be potentially useful for facilitating the design or search of dopamine subtype selective ligands.

3.3.6 Virtual screening performance of the two-step binary relevance SVM method in searching estrogen receptor subtype selective ligands

The VS performance of the SVM models for each ER subtype developed by the 10 sets of randomly assembled training and testing datasets is provided in **Supplementary Table S2**. The sensitivity, specificity, overall accuracy and the Matthews correlation coefficients of these SVM models in classifying ER subtype ligands and non-ligands are in the range of 92.9%-97.6%, 99.7%-99.9%, 99.7%-99.9%, and 0.84-0.92 respectively, which are very similar to those of the dopamine receptor subtype. Moreover, as shown in **Table 3-13 and 3-14**, the performance of 2SBR-SVM in identifying ERα selective ligands (85.0%), ERβ selective ligands (80.0%), ERα and ERβ multi-subtype ligands (69.8%), and in screening Pubchem, MDDR and ChEMBLdb compounds (virtual hit rates 0.0094%-0.0104%, 0.056%-0.064%, and 0.033%-0.034%) is at very similar levels as those of the dopamine receptor subtype. Therefore, our 2BR-SVM method is likely applicable to different receptor-ligand systems.

Table 3-13 The performance of our new method 2SBR-SVM and that of previously used methods Combi-SVM, ML-kNN and RAkEL-DT in predicting estrogen receptor subtype selective and multi-subtype ligands.

Type of estrogen receptor ligands	**Number of ligands**	**Percent of these ligands correctly identified by method**			
		Combi-SVM	ML-kNN	RAkEL-DT	2SBR-SVM
ERα selective ligands	40	55.00%	40.00%	52.50%	85.00%
ERβ selective ligands	55	60.00%	54.55%	58.18%	80.00%
ERα and ERβ multi-subtype ligands	63	63.49%	44.44%	49.20%	69.84%

Table 3-14 Virtual screening performance of our new method 2SBR-SVM and that of our previously used method Combi-SVM in scanning 13.56 million Pubchem compounds, 168,016 MDDR compounds and 657,736 ChEMBLdb compounds. For comparison, the results of single label SVM, which identify putative subtype binding ligands regardless of their possible binding to another subtypes, are also included.

Estrogen receptor subtype	Method	Number and Percent of the 13.56M PubChem Compounds Identified as subtype selective ligands	Number and Percent of the 168,016 MDDR Compounds Identified as subtype selective ligands	Number and Percent of the 657,736 ChemBL Compounds Identified as subtype selective ligands
ERalpha	SVM (Single Label)	19508(0.1439%)	1395(0.8303%)	2689(0.4088%)
	Combi-SVM	9570(0.0706%)	1075(0.6398%)	1931(0.2936%)
	2SBR-SVM	1279(0.0094%)	107(0.0637%)	221(0.0336%)
ERbeta	SVM (Single Label)	20067(0.1480%)	1167(0.6946%)	3017(0.4587%)
	Combi-SVM	10756(0.0793%)	768(0.4571%)	1562(0.2375%)
	2SBR-SVM	1364(0.0101%)	94(0.0559%)	215(0.0327%)

3.4 Conclusion

Virtual screening methods have been increasingly explored for facilitating the discovery of target selective drugs for enhanced therapeutics and reduced side effects. Our study further suggested that the two-step target binding and selectivity support vector machines virtual screening tools developed from protein subtype

ligands with unspecified subtype selectivity are capable of identifying protein subtype selective ligands at good yields, subtype selectivity and low false-hit rates in screening large chemical libraries. Our method may be combined with other virtual screening methods [37, 237-242] to facilitate more effective and efficient search of novel subtype selective drug leads from larger chemical libraries. The capability of virtual screening tools can be further enhanced by the incorporation of the knowledge of existing and newly discovered subtype selective [91, 93] and multi-subtype [204, 205] ligands, and by the further improvement of virtual screening algorithms and parameters [206, 243-248].

Chapter 4 Virtual Screening Prediction of IKK beta Inhibitors from Large Compound Libraries by Support Vector Machines

Summary

The activation of the nuclear factor kappa B (NF-κB) signaling pathway which converge on a serine/threonine kinase plays a key role in the activation of NF-κB: the I kappa B kinase β (IKKβ). Therefore, IKKβ is considered an interesting target for combating inflammation and cancer. However, virtual screening study of potential IKKβ has not been used in large libraries. In this chapter, machine learning based virtual screening models were built to predict the potential IKKβ inhibitors.

4.1 Introduction

The cytotoxicity of chemotherapeutic agents is attributed to apoptosis. Acquired resistance to the effect of chemotherapy has become a serious impediment to effective cancer therapy. The cytotoxic treatments of cancer share a common that their activation of the transcription factor nuclear factor-κB, which regulates cell survival. Activation of the NF-κB signaling pathway by various stimuli, of which tumor necrosis factor-α (TNF-α) is probably the most extensively studied, induces the transcription of proinflammatory target genes via the activation of a complex intracellular signaling cascade, involving among others TNF receptor-associated factors (TRAFs) 2, 5, and 6, I kappa B kinases

(IKKs), and inhibitory κB proteins (IκB). The phosphorylation and degradation of IkB have received significant attention as key steps for the regulation of NF-kB complexes [249]. IKKβ forming the IKK complex together with IKKα and the regulatory domain NEMO (IKKγ) is primarily contributing to the phosphorylation of the inhibitor of NF-κB protein, IκB-α, at residues Ser32 and Ser36.[250] As inhibition of IKK-β is disabling NF-κB activation, the development of small molecule IKKβ inhibitors as potential anti-inflammatory and chemosensitizing agents is promising. IKK beta has been a prime target for the development of NF-kB signaling inhibitors [251-253]. In this work, the machine learning method SVM, k-NN and PNN are used to build prediction and virtual screening models of IKKβ inhibitors. There VS performances are compared in screening large libraries.

4.2 Methods

4.2.1 Data collection of IKK beta inhibitors

A total of 907 of IKK beta inhibitors were collected from literatures and Chembldb (http://www.ebi.ac.uk/chembldb/index.php) and BindingDB (http://www.bindingdb.org/bind/index.jsp).The inhibitors were deposited in *sdf* format files which contain the 3D structures and chemical properties.

Small number of non-inhibitors has been reported. In order or improve the overall representative of non-inhibitors, putative non-inhibitors were generated by using our method of generating putative inactive compounds [206, 254]. This method requires no knowledge of known inactive and active compounds of other target classes, which enables more expanded coverage of the “non-inhibitor” chemical space. Although the yet to be discovered inhibitors are likely distributed

in some of these noninhibitor families, a substantial percentage of these inhibitors are expected to be identified as inhibitors rather than non-inhibitors, even though representatives of their families are putatively assigned as non-inhibitors[206]. The 13.56M PubChem and 168K MDDR compounds were grouped into 8 423 compound families by clustering them in the chemical space defined by their molecular descriptors [255, 256].

The collected IKK beta inhibitors were clustered into 331 families. Because of the extensive efforts in searching kinase inhibitors from known compound libraries, the number of undiscovered IKK beta inhibitor families in PubChem and MDDR databases is expected to be relatively small, most likely no more than several hundred families. The ratio of the discovered and undiscovered inhibitor families (hundreds) and the families that contain no known inhibitor of each kinase (8 423 based on the current versions of PubChem and MDDR) is expected to be <15%. Therefore, a putative noninhibitor training data set can be generated by extracting a few representative compounds from each of those families that contain no known inhibitor, with a maximum possible “wrong” classification rate of <15%, even when all of the undiscovered inhibitors are misplaced into the noninhibitor class. The noise level generated by up to 15% wrong negative family representation is expected to be substantially smaller than that of the maximum 50% false-negative noise level tolerated by SVM[169]. Based on earlier studies [206, 254] and this work, it is expected that a substantial percentage of the undiscovered inhibitors in the putative noninhibitor families can be classified as inhibitor despite that their family representatives are placed into the noninhibitor training sets.

4.2.2 Molecular Descriptors

Molecular descriptors are quantitative representations of structural and physicochemical features of molecules, which have been extensively used in deriving structure-activity relationships[257, 258], quantitative structure activity relationships[259, 260], and VS tools[173, 174, 261-265]. All of the 98 1D and 2D descriptors available from our software[266] were used in this work so as to optimally represent the chemical space covered by the 13.56M PubChem and 168K MDDR compounds. These descriptors and the relevant references are given in Table 2, which include 18 descriptors in the class of simple molecular properties, three descriptors in the class of chemical properties, 42 descriptors in the class of electro-topological state, and 35 descriptors in the class of molecular connectivity and shape. Descriptors in the first three classes are non-redundant. Some descriptors in the fourth class have some degree of overlap in describing the topological features in spite of their differences in mathematical expression. These descriptors include the Schultz molecular topological index, the Gutman molecular topological index, the Wiener index, the Harary index, and the gravitational topological index. The partial overlap in the topological descriptors is not expected to be a serious problem for SVM classification because SVM is less penalized by descriptor redundancy[166, 167].

4.2.3 Support Vector Machines (SVM)

SVM, illustrated in Figure 1, is a supervised ML method based on the structural risk minimization principle for minimizing both training and generalization error [164]. There are linear and nonlinear SVMs. 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:

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

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

where y_i is the class index, **w** is a vector normal to the hyperplane, $|b|/\|\mathbf{w}\|$ is the perpendicular distance from the hyperplane to the origin and $\|\mathbf{w}\|^2$ is the Euclidean norm of **w**. Base on **w** and *b*, a given vector ***x*** can be classified by *f(**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[168-175], 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[176-178]. 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 $f(\boldsymbol{x})$ value indicates that the vector $\boldsymbol{x}$ is an inhibitor or non-inhibitor respectively.

In developing our SVM VS tool, a hard margin c=100,000 was used, and the σ values were found to be 1.4. In 5-fold cross validation studies, the inhibitor and non-inhibitor prediction accuracies are given by sensitivity and specificity respectively. Prediction accuracies have also been frequently measured by overall prediction accuracy (*Q*) and Matthews correlation coefficient (*C*). In the large database screening tests, the yield and false-hit rate are given respectively. The detailed performance evaluation is described in Chapter 2 (section 2.4).

K-NN and PNN methods are described in Chapter 2 (section 2.3).

4.3 Results

4.3.1 Performance of SVM identification of IKK beta inhibitors based on 5-fold cross validation test

The 5-fold cross validation test results of SVM in identifying IKK beta inhibitors and putative non-inhibitors are given in **Table 4-1**. The accuracies for predicting inhibitors and non-inhibitors are 84.07%~94.48% and 99.92%~99.98% respectively. The overall prediction accuracy *Q* and Matthews correlation coefficient *C* are 99.58%~99.88% and 0.796~0.911 respectively.

Table 4-1 Performance of support vector machines for identifying IKK beta inhibitors and non-inhibitors evaluated by 5-fold cross validation study

Cross -Validation	IKK beta inhibitors				IKK beta non-inhibitors				Q (%)	C
	Number of training/ testing inhibitors	TP	FN	SE(%)	Number of training/ testing non-inhibitors	TN	FP	SP(%)		
1	689/181	168	13	92.82%	51782/12946	12943	3	99.98%	99.88%	0.911
2	688/182	153	29	84.07%	51784/12947	12941	6	99.98%	99.58%	0.796
3	690/182	165	17	90.66%	51780/12947	12937	10	99.92%	99.79%	0.853
4	692/181	171	10	94.48%	51782/12947	12936	11	99.92%	99.84%	0.886
5	692/181	166	15	91.71%	51784/12947	12937	10	99.92%	99.81%	0.863
average				**90.75%**				**99.94%**	**99.78%**	**0.862**

4.3.2 Virtual screening performance of SVM in searching IKKb inhibitors from large compound libraries

SVM VS tool for searching IKKβ inhibitors from large were developed by using IKKβ kinases reported before 2009 as described in the methods section. The VS performance of SVM in identifying IKKβ inhibitors reported since 2009 and in searching MDDR and PubChem databases is summarised in **Table 4-2**. The yield in searching IKKβ inhibitors reported since 2009 is 66.96%, which is comparable to the reported 57.86%~71.4% yields of various VS tools[240].

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 262 and 0.16% respectively. The numbers of virtual-hits and virtual-hit rates in screening 13.56M PubChem compounds are 7513 and 0.06% respectively.

Table 4-2 Virtual screening performance of support vector machines for identifying IKK beta inhibitors from large compound libraries.

Method	**Inhibitors in Training Set**		**Inhibitors in Testing Set**			**Virtual screening performance**		
	Number of Inhibitors	**Number of Chemical Families Covered by Inhibitors**	**Number of Inhibitors**	**Number of Chemical Families Covered by Inhibitors**	**Percent of Inhibitors in Chemical Families Covered by Inhibitors in Training Set**	**Yield**	**Number and Percent of 13.56M PubChem Compounds Identified as Inhibitors**	**Number and Percent of the 168K MDDR Compounds Identified as Inhibitors**
Support Vector Machines	729	251	112	77	15.58%	66.96%	7513 (0.06%)	262 (0.16%)
Tarnimoto Similarity						27.67%	97,675(0.72%)	4,042(2.45%)
K Nearest Neighbour						57.86%	79,855(0.58%)	1,702(1.0%)
Probabilistic Neural Network						71.40%	84,567(0.62%)	1,821(1.08%)

4.3.3 Comparison of Performance of SVM-based and other VS methods

To evaluate the level of performance of SVM and whether the performance is due to the SVM classification models or to the molecular descriptors used, SVM results were compared with those of three other VS methods based on the same molecular descriptors, training dataset of IKK beta inhibitors reported before 2009, and the testing dataset of IKK beta inhibitors reported since 2009, 168K MDDR and 13.56M PubChem compounds. The three other VS methods include two similarity-based methods, Tanimoto-based similarity searching and kNN methods, and an alternative machine learning method PNN. As shown in **Table 4-2**, the yield of the Tanimoto-based similarity searching, kNN and PNN methods are 27.67%, 57.86%, and 71.4% respectively. Compared to these results, the yield of SVM is smaller than PNN but higher than other two similarity-based VS methods. These suggest that SVM performance is due primarily to the SVM classification models rather than the molecular descriptors used. The false-hit rate of SVM method are 0.06% to 0.16% for PubChem and MDDR libraries, which are considerably small than other methods. Our results are consistent with the report that SVM shows mostly good performances both on classification and regression tasks, but other classification and regression methods proved to be very competitive[267].

4.4 Conclusion Remarks

SVM shows substantial capability in identifying IKK beta inhibitors at comparable yield and in many cases substantially lower false-hit rate than those of typical VS tools reported in the literatures and evaluated in this work. 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. 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 for facilitating the discovery of IKK beta inhibitors and other active compounds.

Chapter 5 Analysis of bypass signaling in EGFR pathway and profiling of bypass genes for predicting response to anticancer EGFR tyrosine kinase inhibitors

Summary

Some drugs such as anticancer EGFR tyrosine kinase inhibitors elicit markedly different clinical response rates due to differences in drug bypass signaling as well as genetic variations of drug target and downstream drug-resistant genes. The profiles of these bypass signaling are expected to be useful for improved drug response prediction, which have not been systematically explored. In this work, we searched and analyzed 16 literature-reported EGFR tyrosine kinase inhibitor bypass signaling routes in EGFR pathway, which include 5 compensatory routes of EGFR transactivation by another receptor, and 11 alternative routes activated by another receptor. These 16 routes are reportedly regulated by 11 bypass genes. Their expression profiles together with the mutational, amplification and expression profiles of EGFR and 4 downstream drug-resistant genes were used as new sets of biomarkers for identifying 53 NSCLC cell-lines sensitive or resistant to EGFR tyrosine kinase inhibitors gefitinib, erlotinib, and lapatinib. The collective profiles of all 16 genes distinguish sensitive and resistant cell-lines are better than those of individual genes and the combined EGFR and downstream drug resistant genes, and their derived cell-line response rates are consistent with the reported clinical response rates of the three drugs. The usefulness of cell-line data for drug response studies was further analyzed by comparing the expression

profiles of EGFR and bypass genes in NSCLC cell-lines and patient samples, and by using a machine learning feature selection method for selecting drug response biomarkers. Our study suggested that the profiles of drug bypass signaling are highly useful for improved drug response prediction.

5.1 Introduction

The highly successful anticancer kinase inhibitor drugs typically elicit markedly different anticancer clinical response rates [268, 269]. For instance, the reported clinical response rates of EGFR tyrosine kinase inhibitors (EGFR-I) gefitinib and erlotinib are 19.9% and 8.9% respectively for the treatment of non-small cell lung cancer (NSCLC) [268, 269]. These differences have been linked to activating mutations, amplification and over expression of EGFR [269-272], and such genetic variations of EGFR downstream drug-resistant genes as activating mutations of RAS, BRAF, PIK3CA and AKT [269, 273-275], and loss-of-function of PTEN (including PTEN loss and inactivating mutations) [276-278]. The genetic and expression profiles of EGFR and downstream drug-resistant genes have been individually and collectively explored as biomarkers for predicting clinical response to EGFR-I [279-284]. Apart from these genes, drug response can also be significantly altered by compensatory, alternative and redundant signaling that bypass drug actions [282, 285]. Understanding the mechanism of these bypass signaling events is useful not only for discovering multi-target drugs and drug combinations[286-290] but also for discovering new biomarkers that in combination with existing biomarkers improves drug response prediction [282, 285, 290, 291].

Drug response biomarkers besides drug target and downstream drug-resistant genes can be derived by profiling the expression patterns of thousands of genes [292-294] and proteins [295]. While genome-scale gene and protein expression profiling is capable of predicting disease [104, 106, 296] and treatment [297] outcomes at high accuracy levels, data noise and other factors may in some cases affect the stability of derived biomarkers and the quality of identified disease or drug response genes [298-302]. Therefore, biomarker discovery from systems perspective of drug response signaling is expected to complement existing biomarker discovery methods by correlating drug response to the collective profiles of the drug target, downstream resistant genes, and bypass signaling.

One of the successfully used experimental strategies for drug response mechanism investigations and biomarker discovery involves the comparative analysis of drug-resistant and drug-sensitive cancer-derived cell-line models [303], which has been used for implicating the compensatory PI3K/Akt/mTor activation as a mechanism of acquired resistance to imatinib in chronic myeloid leukemia (CML) [304], MET amplification as a mechanism of acquired resistance to EGFR-I therapy in NSCLC [305], and CRAF overexpression as a potential mechanism of acquired resistance to BRAF inhibitor therapy in melanomas [306].

Drug resistance related genetic mutations[307-310], chromosomal changes [311], gene expression [312-318], gene amplification [319, 320], and combinations of these profiles [321-324] have been actively explored as drug response biomarkers. Some of the identified biomarkers are related to the genes that regulate drug bypass signalling [293, 312, 319, 320, 325, 326]. These studies demonstrate the effectiveness of collective analysis of genetic and proteomic profiles of the drug targets and bypass signaling in predicting drug response. Moreover, some of the

relevant studies are based on cell-line models [312, 314, 319, 326], suggesting that cell-line models have some level of usefulness in facilitating drug response studies.

In predicting drug response, the presence of known drug resistance mutations is used for indicating drug resistance [307, 309, 310]. Drug response can also be predicted from the expression profiles of the genes involved in the drug response regulations [312, 319, 320, 325, 326], with the up-regulation of the genes promoting drug bypass signalling [325, 326] and down-regulation of the genes [325] against drug actions [327] used for suggesting drug resistance. Previously unknown drug response biomarkers can be predicted from gene expression profiles by using the principal components analysis feature selection method [312], weighted voting classification feature selection method [314], hierarchical clustering feature selection method [313], differentially expressed genes method,[315, 328] and machine learning feature selection methods [295, 316]. The amplification of genes involved in drug bypass signalling is also linked to drug resistance.

While cell-lines have been widely used for drug response studies [312, 314, 319, 326] as well as in basic and translational biomedical research and drug discovery campaigns [303, 329, 330], the limited number of cell-lines may not adequately capture the heterogeneous nature of real tumors [303]. The genome instability [331, 332], plasticity [333], and microenvironment [334, 335] of cell-lines may differ significantly from those of real tumors, which likely affect the quality of drug response studies [334]. For instance, the potency of trastuzumab has been found to change significantly when tumor cells are grown in 2D culture or 3D matrix [336] and when cells are plated on different extracellular matrices [337].

In this work, we searched and analysed literature reported compensatory,

alternative and redundant signaling pathways that have been experimentally found to contribute to the resistance of EGFR-I, and identified the corresponding bypass genes experimentally confirmed to play key roles in regulating these bypass signaling. The level of contribution of each bypass signaling pathway against EGFR-I is determined by the genetic and expression profiles of the corresponding bypass genes [338]. These profiles differ from individual to individual and thus are expected to contribute to the individual variations in drug response. To determine the level of usefulness of the profiles of the reported bypass genes in drug response prediction, we retrospectively analysed the expression profiles of 11 bypass genes (HER2, HER3, IGF1R, c-MET, PDGFR, FGFR, VEGFR2, Integrin β1, MDGI , IL-6 and Cox2) together with the mutational, amplification and gene expression profiles of EGFR and 4 downstream drug-resistant genes (RAS, BRAF, PIK3CA, and PTEN) in 53 NSCLC cell-lines sensitive or resistant to EGFR-I gefitinib, erlotinib, and lapatinib (**Table 5-1** and **5-2**). The presence of drug resistance mutations [307, 309, 310], PTEN loss of function [339], enhanced accumulation of internalized EGFR [340], and up-regulation [325, 326] and amplification of the bypass genes were used for predicting a cell-line to be resistant to an EGFR-I. The drug response prediction performance of the collective profiles of the 16 genes in these cell-lines is compared to those of individual profiles and the combined profiles of EGFR and 4 downstream drug-resistant genes. The overall cell-line response rates of the 3 drugs were also compared with the reported clinical response rates in NSCLC patients to determine to what extent cell-line derived response rates agree with the clinical response rates.

Table 5-1 The bypass genes, regulated bypass signaling or regulatory genes, and the relevant bypass mechanisms in the treatment of NSCLC.

Bypass Gene	Regulated Bypass Signaling	Bypass Mechanism
HER2	Compensatory signaling via EGFR-HER2 transactivation, and subsequent activation of RAS and AKT pathways	EGFR inhibition upregulated HER2 and induced compensatory EGFR-HER2 heterodimerisation to promote alternative signaling [341, 342]
	Alternative signaling via HER2-HER3 and HER2-HER4 transactivation, and subsequent activation of RAS and AKT pathways	EGFR inhibition upregulated HER2 and induced HER2-HER3, HER2-HER4 heterodimerisation to promote alternative signaling[341, 342]
HER3	Compensatory signaling via EGFR-HER3 transactivation, and subsequent activation of ATK pathway	EGFR inhibition elevated HER3 and subsequently induced compensatory transactivation of HER3 signaling [285]
	Alternative signaling via HER2-HER3 transactivation, and subsequent activation of RAS and ATK pathways	EGFR inhibition upregulated HER2 and induced HER2-HER3 heterodimerisation to promote alternative signaling[342, 343]
	Alternative signaling via HER3 autophosphorylation, and subsequent activation of ATK pathway	HER3 may autophosphorylate to produce weak kinase activity that may contribute to the resistance of EGFR inhibitor[344]
	Alternative signaling via PDGFR-HER3 transactivation, and subsequent activation of RAS and ATK pathway	EGFR inhibition countered by PDGFR transactivation of HER3 signaling[286]
IGF1R	Compensatory signaling via EGFR-IGF1R transactivation, and subsequent activation of RAS and ATK pathway	EGFR inhibition upregulated IGF1R and induced EGFR-IGF1R heterodimerization and activation of IGFR signaling[345]
	Alternative signaling via IGF1R activation, and subsequent activation of RAS and ATK pathway	EGFR inhibition reduced IGF-binding protein IGFBP-3 and IGFBP-4 to derepresses IGFIR signaling[346]
c-MET	Compensatory signaling via EGFR-MET transactivation, and subsequent activation of RAS and	EGFR inhibition countered by focal amplification of MET that physically interacts with EGFR to promote transactivation[286, 347], Met activation

	ATK pathway	in NSCLC is associated with de novo resistance to EGFR inhibitors and the development of metastasis[348]
	Alternative signaling via MET-HER3 transactivation, and subsequent activation of RAS and AkT pathways	EGFR inhibition countered by focal amplification of MET that transactivates HER3 to drive HER3-dependent activation of PI3K[305]
	Alternative signaling via MET-HER2 transactivation, and subsequent activation of RAS and AkT pathways	EGFR inhibition countered by focal amplification of MET that physically interacts with HER2 to promote alternative signaling[286, 347]
	Alternative signaling via HGF-induced MET activation, which subsequently activate MAPK and AKT pathways independent of EGFR and HER3	HGF-induced MET activation re-stimulated the MAPK and AKT pathways independent of EGFR and HER3 and restored cell proliferation, which is a novel mechanism of cetuximab resistance in CRC. Inhibition of the HGF-MET pathway may improve response to EGFR inhibitors in CRC[349]
PDGFR	Alternative signaling via PDGFR-HER3 transactivation, and subsequent activation of RAS and AkT pathways	EGFR inhibition countered by PDGFR transactivation of HER3 signaling[286]
	Alternative signaling via PDGFR autophosphorylation, and subsequent activation of RAS and AkT pathways	PDGF, PDGFR are expressed in certain NSCLC cell-lines, EGFR inhibition induced PDGFR autophosphorylation[350]
FGFR	Alternative signaling via FGF-FGFR autocrine pathway, and subsequent activation of RAS and AkT pathways	FGFR contributed to EGFR inhibitor resistance via alternative signaling[350], an FGF-FGFR autocrine pathway dominates in some NSCLC cell-lines to promote the switch to FGFR signaling[351]
VEGFR2	Alternative signaling via VEGFR2 pathway, and subsequent activation of RAS and AkT pathways	EGFR inhibition shifts tumor population towards a less EGFR-dependent and more VEGF-dependent phynotype, combined blockade of VEGFR and EGFR pathways can abrogate resistance to EGFR inhibitors[352]
Integrin β1	Compensatory signaling via EGFR-Integrin beta1 transactivation, and subsequent activation of RAS and AkT pathways	Integrin beta1 over-expression associates with resistance to gefitinib in NSCLC (21053345), it associates with EGFR, c-SRC and P130 to activate EGFR[353, 354]

	Alternative signaling via integrin beta1 recruitment of FAK and a PP2-sensitive kinase to activate AkT pathway	Integrin β1 over-expression associates with resistance to gefitinib in NSCLC[355], it activates AkT pathway by recruiting either FAK or an upstream PP2-sensitive non SRC tyrosine kinase to activate PI3K[356]
MDGI	Compensatory signaling via enhanced accumulation of internalized EGFR, and enhanced activation of RAS and AkT pathways	MDGI regulated EGFR subcellular localization, MDGI over-expression increased intracellular accumulation of EGFR and may be a biomarker for responsiveness to anti-EGFR antibody therapy[340]
IL-6	Alternative signaling via IL-6 activation of MEK and JAK/STAT	IL-6 is upregulated in Erlotinib-resistant cells and required for their survival, and the up-regulation is mediated by TGF-β signaling, IL-6 activated gp130/JAK/STAT pathway to decrease sensitivity to erlotinib[357], EGFR can activate JAK/STAT via Mek, elevated IL can activate JAK/STAT and Mek to substitute EGFR activation of Mek and STAT[358]
Cox2	Alternative signaling by PGE2 mediated activation of PKC-MEK-ERK pathway and Gβγ-PI3K pathway	Cox2 over-expression caused resistance to Gefitinib and Erlotinib inhibition of Erk[359], Cox2 activated Erk via PGE2–EP receptors-PKC-Ras-Mek, Cox2 activated PI3K via PGE2-EP receptors-Gβγ-PI3K, Cox2 also activated EGFR via PGE2-EP receptors – Src – TGFα –EGFR and PGE2-EP receptors – Amphelegulin-EGFR[360]

Table 5-2 The downstream genes, regulated bypass signaling or regulatory genes, and the relevant bypass mechanisms in the treatment of NSCLC.

Drug Resistant Downstream Gene	Bypass Signaling	Resistance Mechanism
KRAS	Compensatory signaling via EGFR-independent activation of KRAS	KRAS activating mutation mediated EGFR-independent signaling and contributed to EGFR inhibitor resistance[361, 362]
PTEN	Compensatory signaling via enhanced activation of AKT pathway to reduce the level of dependence on EGFR	PTEN loss or inactivating mutation contributed to EGFR inhibitor resistance by activation of Akt and EGFR[278, 280], PTEN-associated resistance to EGFR inhibitors is phenocopied by expression of dominant negative Cbl and can be overcome by more complete EGFR inhibition[363]
PIK3CA	Compensatory signaling via EGFR-independent activation of AKT pathway	PIK3CA activating mutation mediated EGFR-independent AKT signaling and contributed to EGFR inhibitor resistance[275]
AKT		AKT activating mutation mediated EGFR-independent AKT signaling and could lead to resistance against EGFR inhibitor[274]

For non-EGFR addicted tumor cells, the addicted oncogene is the main signaling rather than EGFR-I bypass signaling, and the appropriate therapeutic approach is to target the addicted oncogene instead of EGFR [364]. In this work, these non-EGFR addicted tumor cells were not distinguished from the EGFR addicted cells because they are nonetheless resistant to EGFR tyrosine kinase inhibitors. Large percentage of NSCLC patients contains wild-type EGFR [270, 365]. Although lung cancer patients with wild-type EGFR are less sensitive to the EGFR-Is [366], some EGFR-Is such as erlotinib has been approved as a second/third-line drug for unselected NSCLC based on clinical trial results [367]. Hence, cell-lines with wild-type EGFR were included in this study.

Gefitinib and erlotinib are EGFR-Is approved for lung and pancreatic cancers, and lapatinib is a multi-target EGFR and HER2 inhibitor approved for breast cancer and tested for lung, prostate and liver cancers [271, 288, 368-370]. These drugs were evaluated because of their clinical relevance, knowledge of drug-resistance mechanisms, and availability of drug response, genetic and gene expression data for statistically significant number of cell-lines. The specific genetic data include drug sensitizing mutations and copy number variations in EGFR, activating mutations in RAS, BRAF, PIK3CA and inactivating mutations in PTEN directly contributing to EGFR-I resistance in significant percentage of patients (>2%) [271]. The gene expression data include microarray gene expression data of EGFR, PTEN, and 11 bypass genes directly contributing to EGFR-I resistance [285].

To evaluate the relevance and limitations of cell-line data for drug response studies, we compared the distribution of the known drug resistant mutations and the up-regulated EGFR and bypass genes in our studied NSCLC cell-lines and those in

the real NSCLC patient samples to determine to what extent the cell-line profiles are close to those of real patients. Moreover, the usefulness of cell-line data was further tested by using our support vector machines feature selection method [301] to select EGFR-I response biomarkers and compare them with the known bypass genes and the published EGFR-I response biomarkers derived from NSCLC patient samples [371] and cell-lines [293, 328] by using the differential expression method (genes with the most differential expression in drug resistant and sensitive samples are selected as biomarkers).

5.2 METHODS

5.2.1 EGFR pathway and drug bypass signaling data collection and analysis

We searched the literatures to find experimentally determined bypass signaling and the corresponding bypass genes in response to EGFR inhibition by using keyword combinations of "EGFR", "inhibitor", "resistance", "resistant", "bypass", "insensitive", "sensitivity", and "mechanism". **Table 5-1** and **5-2** summarizes the 11 bypass-genes, 4 downstream drug resistance genes and the regulator of enhanced accumulation of internalized EGFR, the corresponding bypass and resistance mechanisms. The map of the major signaling pathways of EGFR and downstream effectors relevant to cancers [341, 372-374] was shown in **Figure 5-1**. Keyword combinations of "EGFR", "pathway", "signaling" were used to search the literatures that describe the pathway map, and the keyword combinations of protein name, alternative name, "interaction/interacting", "binding/bind",

"regulation/regulated" were used to search relevant protein-protein interactions and regulations from Pubmed [375]. We further searched from the literatures for the activating and resistant mutations of EGFR [271, 376, 377] and downstream drug resistant genes [339, 376, 378-380]. The amino acid sequences of EGFR and 4 downstream drug-resistant genes were from the Swissprot database of UniprotKB [381].

5.2.2 NSCLC cell-lines with EGFR tyrosine kinase inhibitor sensitivity data

We identified from literatures [373, 382, 383] NSCLC cell-lines with available sensitivity data for gefitinib, erlotinib, and lapatinib (**Table 5-3 and 5-4**). Overall, 46 NSCLC cell-lines with sensitivity data for one or more drugs were collected. A cell-line was considered to be sensitive to a drug if the drug inhibits it at IC50≤1μM,[384] otherwise it was considered as resistant to the drug. The genetic and microarray gene expression data for 53 NSCLC cell-lines were obtained from the published literatures, and COSMIC [385] and GEO [386] databases. We further identified from GEO database the microarray gene expression data for 6 lung cell-lines of healthy people respectively. The relevant data and literature sources for these cell-lines are summarized in the **Table 5-5**. These expression data were processed by using RMA normalization method [387].

Table 5-3 Clinicopathological features of NSCLC cell-lines used in this study. The available gene expression data, EGFR amplification status, and drug sensitivity data for gefitinib, erlotinib, and lapatinib are included together with the relevant references.

Cell-line	Histological Type	Histological Subtype *	Source *	Gene Expression Sample ID at GEO Database([388]	EGFR Amplification (gene copy number >3)[280]	EGFR Amplification (gene copy number >4)[280]	Mutated Gene/Genes[389, 390]	Sensitivity Data		
								Gefitinib[391, 392]	Erlotinib[280, 392]	Lapatinib[280]
A427	NSCLC	NS	PT	NA			KRAS		R	R
A549	NSCLC	NS	PT	GSM108799			KRAS	R	R	R
Calu1	NSCLC	EC	PE	GSM108801			KRAS	R	R	R
Calu3	NSCLC	AD	PE	GSM108803				S		S
Calu6	NSCLC	APC	PT	GSM108805			KRAS	R	R	R
Colo699	NSCLC	AD	PF	NA	Y				R	R
DV90	NSCLC	AD	PE	NA			KRAS		R	R
EKVX	NSCLC	AD	PT	NA					R	R
H1155	NSCLC	LCC	PT	NA	NA	NA	KRAS,PTEN	R	R	
H1299	NSCLC	LCC	LN	GSM108807			NRAS	R	R	R
H1355	NSCLC	AD	PT	GSM108809			KRAS, BRAF	R	R	R
H1395	NSCLC	AD	PT	GSM108811			BRAF	R	R	R
H1437	NSCLC	AD	PT	GSM108813				R	R	R
H1563	NSCLC	AD	PT	NA			PIK3CA		R	R
H1568	NSCLC	AD	PT	NA	Y	Y			R	R
H157	NSCLC	SQ	PT	GSM108815			KRAS,PTEN	R	R	R
H1648	NSCLC	AD	LN	GSM108817				R	R	S
H1650	NSCLC	AD	PE	GSM108819	Y		EGFR	R	R	R
H1666	NSCLC	AD	PE	GSM108821			BRAF	R	R	S

H1734	NSCLC	AD	PT	NA	Y		KRAS		R	R
H1755	NSCLC	AD	Live	NA			BRAF		R	R
H1770	NSCLC	NE	LN	GSM108825				R	R	
H1781	NSCLC	AD	PE	NA			ERBB2	R	R	R
H1792	NSCLC	AD	PE	GSM171848	Y		KRAS		R	R
H1819	NSCLC	AD	LN	GSM108827	Y			R	R	S
H1838	NSCLC	AD	PT	NA	Y	Y			R	R
H1915	NSCLC	SCC	Brain	NA					R	R
H1944	NSCLC	AD	ST	NA			KRAS		R	R
H1975	NSCLC	AD	PT	GSM108829	Y		EGFR	R	R	R
H1993	NSCLC	AD	LN	GSM108831				R	R	R
H2009	NSCLC	AD	LN	GSM108833			KRAS	R	R	R
H2030	NSCLC	AD	LN	NA			KRAS		R	R
H2052	NSCLC	MT	PE	GSM171854					R	R
H2077	NSCLC	AD	PT	NA					R	R
H2087	NSCLC	AD	LN	GSM108835			BRAF, NRAS	R	R	R
H2110	NSCLC	NS	PE	NA					R	R
H2122	NSCLC	AD	PE	GSM108837			KRAS	R	R	R
H2126	NSCLC	LCC	PE	GSM108839				R	R	R
H2172	NSCLC	NS	PT	NA					R	R
H2228	NSCLC	AD	PT	NA					R	R
H23	NSCLC	AD	PT	GSM171868			KRAS, PTEN		R	R
H2347	NSCLC	AD	PT	GSM108841			NRAS	R	R	R
H2444	NSCLC	NS	PT	NA	Y		KRAS		R	R
H28	NSCLC	MT	PE	GSM171870					R	R

H2882	NSCLC	NS	PT	GSM108843				R	R	R
H2887	NSCLC	NS	PT	GSM108845			KRAS	R	R	R
H3122	NSCLC	AD	PT	GSM171874					R	R
H322	NSCLC	AD	PT	GSM171876	Y			R	R	R
H322M	NSCLC	AD	PT	NA					R	S
H3255	NSCLC	AD	PT	GSM108847	Y	Y	EGFR	S	S	S
H358	NSCLC	AD	PT	GSM108849			KRAS	R	R	R
H441	NSCLC	AD	PT	GSM108851			KRAS	R	R	R
H460	NSCLC	LCC	PE	GSM108853			KRAS, PIK3CA	R	R	R
H520	NSCLC	SQ	PT	NA				R	R	R
H522	NSCLC	AD	PT	NA	Y				R	R
H596	NSCLC	AD	PT	NA	Y		PIK3CA		R	R
H647	NSCLC	ADSQ	PE	NA			KRAS		R	R
H661	NSCLC	LC	LN	GSM171884					R	R
H820	NSCLC	AD	LN	GSM108855	Y		EGFR	R	R	R
HCC1171	NSCLC	NS	PT	GSM108857			KRAS	R	R	R
HCC1195	NSCLC	ADSQ	PT	GSM108859	Y		NRAS	R	R	R
HCC1359	NSCLC	SGC	PT	GSM108861				R	R	R
HCC15	NSCLC	SQ	PT	GSM108863			NRAS	R	R	R
HCC1833	NSCLC	AD	PT	GSM171898					R	R
HCC193	NSCLC	AD	PT	GSM108865	Y			R	R	R
HCC2279	NSCLC	AD	PT	GSM108867	Y	Y	EGFR	S	S	R
HCC2429	NSCLC	NS	PT	GSM171900					R	R
HCC2450	NSCLC	SQ	PT	GSM171902			PIK3CA		R	R
HCC2935	NSCLC	AD	PE	GSM108869			EGFR	S	S	S

HCC364	NSCLC	AD	PT	NA			BRAF		R	R
HCC366	NSCLC	ADSQ	PT	GSM108871				R	R	R
HCC4006	NSCLC	AD	PE	GSM108873	Y	Y	EGFR	S	S	S
HCC44	NSCLC	AD	PT	GSM108875			KRAS	R	R	R
HCC461	NSCLC	AD	PT	GSM108877			KRAS	R	R	R
HCC515	NSCLC	AD	PT	GSM108879			KRAS	R	R	R
HCC78	NSCLC	AD	PE	GSM108881				R	R	R
HCC827	NSCLC	AD	PT	GSM108883	Y	Y	EGFR	S	S	S
HCC95	NSCLC	SQ	PE	GSM108885				R	R	R
HOP62	NSCLC	AD	PT	NA			KRAS		R	R
HOP92	NSCLC	AD	PT	NA	Y				R	R
LCLC103H	NSCLC	LCC	PE	NA					R	R
LCLC97TM	NSCLC	LCC	PT	NA			KRAS		R	R
LouNH91	NSCLC	SQ	PT	NA	Y		EGFR		R	R
PC9	NSCLC	AD	PT	NA	Y		EGFR	S	S	R
SKLU1	NSCLC	AD	PT	NA			KRAS		R	R

* Determined from the ATCC (http://www.atcc.org) and DSMZ (http://www.dsmz.de) websites, and references therein.

Abbreviations: AD, lung adenocarcinoma; APC, anaplastic carcinoma; EC, epidermoid carcinoma; LCC, large cell lung cancer; LN, lymph node; MT, mesothelioma; NA: not available; NE, neuroendocrine neoplasm; NS, not specified; NSCLC: non-small cell lung cancer; PE, pleural effusion; PF, pleural fluid; PT, primary tumor; R, resistant; S, sensitive ; SCC, small-cell carcinoma; SGC: spindle and giant cell carcinoma; ST, soft tissue; Y, gene amplified

Table 5-4 Sensitivity data of NSCLC cell-lines treated with gefitinib, erlotinib, and lapatinib.

NSCLC Cell-line	Sensitivity of Cell-line to Gefitinib Inhibition [#]	Reported Potency (IC50) of Gefitinib Inhibition (μM)		Sensitivity of Cell-line to Erlotinib Inhibition [#]	Reported Potency (IC50/ED50) of Erlotinib Inhibition (μM)		Sensitivity of Cell-line to Lapatinib Inhibition [#]	Reported Potency (ED50) of Lapatinib Inhibition (μM)
		Ref [392]	Ref [391]		Ref [392]	Ref [280]		Ref [280]
A427				R		1.24	R	9.4406
A549	R	25		R	60	10	R	10
Calu1	R		41	R		10	R	10
Calu3	S	0.78		-	1.29	0.7	S	0.1679
Calu6	R		34	R		9.65	R	2.7542
Colo699				R		4.26	R	5.8884
DV90				R		3.95	R	1.4125
EKVX				R		10	R	10
H1155	R	183		R	8.63			
H1299	R	26.4		R	41.9	10	R	10
H1355	R	325		R	27	3.31	R	5.6885
H1395	R	71		R	10.5	5.05	R	6.6834
H1437	R	62		R	12.5	10	R	10
H1563				R		10	R	10
H1568				R		1.08	R	2.541
H157	R	115		R	128	10	R	10
H1648	R	36.7		R	34	7.77	S	0.9441
H1650	R	11.7		R	15	2.13	R	3.8905
H1666	R	180		R	13	3.31	S	0.5957
H1734				R		3.79	R	4.3652
H1755				R		7.5	R	10

H1770	R	160		R	111	10		
H1781	R	19		R	44	2.54	R	2.9174
H1792				R		10	R	10
H1819	R	19		R	6.3	3.92	S	0.7328
H1838				R		3.47	R	10
H1915				R		10	R	10
H1944				R		1.83	R	10
H1975	R	25		R	33	10	R	10
H1993	R	17.9		R	5.2	8.06	R	4.3152
H2009	R	33.2		R	25.8	10	R	10
H2030				R		4.95	R	5.0119
H2052				R		8.98	R	10
H2077				R		10	R	10
H2087	R	18.4		R	9.9	10	R	10
H2110				R		4.5	R	2.7861
H2122	R	35		R	76.8	10	R	10
H2126	R	21.4		R	13	10	R	10
H2172				R		10	R	8.9125
H2228				R		10	R	10
H23				R		10	R	5.6234
H2347	R	60		R	5.2	10	R	5.9566
H2444				R		4.22	R	7.6736
H28				R		10	R	1.6032
H2882	R	19.2		R	66	10	R	5.1286
H2887	R	110		R	101	10	R	10
H3122				R		10	R	10

H322	R	120		R	56	2.21	R	2.4831
H322M				R		1.29	S	0.4416
H3255	S	0.089		S	0.129	0.02	S	0.309
H358	R	12.5		R	6.2	1.11	R	1.6032
H441	R	15.7		R	7.1	3.61	R	10
H460	R	16.9		R	72	10	R	3.3113
H520	R	13.6		R		10	R	6.8391
H522				R		5.83	R	8.7096
H596				R		1.2	R	10
H647				R		10	R	10
H661				R		10	R	10
H820	R	3		R	7.1	10	R	10
HCC1171	R	127		R	160	10	R	10
HCC1195	R	27.6		R	175	10	NA	
HCC1359	R	65		R	88	10	R	10
HCC15	R	52		R	100	10	R	10
HCC1833				R		10	R	2.6915
HCC193	R	21.1		R	20.5	10	R	1.7378
HCC2279	S	0.0479		S	0.093	0.01	R	10
HCC2429				R		10	R	5.9566
HCC2450				R		10	R	10
HCC2935	S	0.11		S	0.163	0.07	S	0.2344
HCC364				R		4.19	R	10
HCC366	R	30		R	11	0.99	R	10
HCC4006	S	0.23		S	0.124	0.04	S	0.537
HCC44	R	57.8		R	28	10	R	10

HCC461	R	13.9		R	16	9.04	R	10
HCC515	R	120		R	154	1.85	R	9.5499
HCC78	R	81		R	21.2	10	R	4.1687
HCC827	S	0.04		S	0.0388	0.02	S	0.7943
HCC95	R	24		R	18.4	10	R	3.2359
HOP62				R		10	R	5.4325
HOP92				R		10	R	10
LCLC103H				R		10	R	10
LCLC97TM				R		5.26	R	7.3282
LouNH91				R		3.05	R	5.1286
PC9	S	0.0309		S		0.02	R	1.4962
SKLU1				R		10	R	10

* A cell-line with IC50≤ 1 μmol/L for gefitinib, erlotinib, and lapatinib was considered to be sensitive (S) to a given drug[384], otherwise it was considered as resistant (R) to the drug. - : cell-line with inconsistent sensitivity data, which is not included in this study.

Table 5-5 6 normal Cell-lines from the lung bronchial epithelial tissues obtained from GEO database.

Gene Expression Sample ID of Normal Cell-line at GEO Database	Cell-lines	Source of Cell-lines	Reference
GSM427196	NHBE	Normal human bronchial epithelial cells	Ref [393]
GSM427197	NHBE	Normal human bronchial epithelial cells	
GSM427198	BEAS-2B	Immortalized bronchial epithelial cells	
GSM427199	BEAS-2B	Immortalized bronchial epithelial cells	
GSM427200	1799	Immortalized lung epithelial cells	
GSM427201	1799	Immortalized lung epithelial cells	

5.2.3 Genetic and expression profiling of bypass genes for predicting drug sensitivity of NSCLC cell-lines

The clinical efficacy of gefitinib, erlotinib, and lapatinib against NSCLC are mostly due to their inhibition of the main target EGFR [271, 288]. Resistance to EGFR tyrosine kinase inhibitors primarily arises from resistant mutations and amplification of EGFR, activating mutations of down-stream signaling genes and loss of function of down-stream negative regulators, and compensatory, alternative and redundant signaling genes frequently up-regulated or amplified in resistant patients [285, 305]. Efflux-pumps, primarily responsible for the resistance of chemotherapy drugs [394], are not expected to significantly contribute to the resistance of the evaluated drugs because these drugs are either efflux-pump inhibitors [383, 395, 396].

We retrospectively evaluated the capability of the individual and combinations of the genetic and expression profiles of the main target, downstream drug resistance genes, and bypass genes in **Table 5-1** and **5-2** for predicting the sensitivity of the 53 NSCLC cell-lines to gefitinib (6 sensitive, 38 resistant), erlotinib (7 sensitive, 46 resistant), and lapatinib (8 sensitive, 40 resistant). We evaluated 14 mutation-based, amplification-based, expression-based, and combination methods by calculating the percentages of correctly predicted sensitive and resistant cell-lines. Due to inadequate copy number data, the amplification-based methods exclude the profiles of the bypass and downstream genes, some of which are known to directly contribute to EGFR tyrosine kinase inhibitor resistance [305]. Nonetheless, copy number variation significantly influences gene expression, with 62% of amplified genes showing moderately or highly elevated expression [397]. Thus the effects of

amplification of bypass genes are expected to be partially reflected by the expression profiles.

In mutation-based method M1, a NSCLC cell-line is predicted as sensitive to a drug if EGFR contains no resistance mutation against the given drug [269] and the drug inhibits EGFR at IC50 ≤ 500nM[384] (a stricter condition of IC50 ≤ 100nM gives the same results in all studied cases), otherwise it is predicted as drug-resistant. In mutation-based method M2, a NSCLC cell-line is predicted as sensitive to a drug if the drug inhibits EGFR at IC50 ≤ 500nM,[384] EGFR contains no resistance mutation against the given drug [269], and the un-inhibited KRAS has no activating mutation [362]. In mutation-based method M3, a NSCLC cell-line is predicted as sensitive to a drug if: (I) the drug inhibits EGFR at IC50≤500nM[384] and EGFR in NSCLC cell-line has no resistance mutation against the given drug [269], (II) the un-inhibited KRAS, NRAS, BRAF, PIK3CA in NSCLC cell-line [271] has no activating mutation, (III) there is no PTEN loss or PTEN inactivating mutation in NSCLC [271] cell-line.

The mutation profiles of the relevant genes in each cell-line were generated by comparative sequence analysis with respect to the reported sensitizing, activating or inactivating mutations, which are summarized in **Table 5-6** and **5-7**. PTEN loss was assumed to occur if its microarray gene expression level is ≤ 1/5 of the median level of PTEN in the normal tissue cell-lines [398], based on the comparison of the western-blot staining of a PTEN-deficient cell-line ZR-75-1 with that of a PTEN-normal cancer cell-line MCF-7 [399] (variation of this cut-off from 0 to 1/3 of the median level gives the same results in all studied cases).

Table 5-6 Drug related sensitizing/resistant mutations of EGFR and cancer related activating mutations of EGFR, PIK3CA, RAS, and BRAF, and inactivation mutations of PTEN.

Disease	Type of Mutation	Percentage of 85 NSCLC Cell-lines or 40 Breast Cancer Cell-lines with This Type of Mutation	Specific Mutations (Number of NSCLC or Breast Cancer Cell-lines with This Mutation)
NSCLC	Gefitinib , erlotinib , and lapatinib sensitizing mutation of EGFR[271]	11.7%	E746_A750del (4) / E746_A750del, T751A(1) / E746_T751del, I ins(1) / L747_E749del, A750P(1) / L747_S752del, P753S(1) / L858R(2)
	Gefitinib , erlotinib , and lapatinib resistant mutation of EGFR[271]	2.4%	T790M (2)
	Gefitinib and erlotinib resistant mutation of HER2[400]	1.2%	G776VC (1)
	Activating mutation of KRAS[378]	32.9%	G12A (1) / G12C (9) / G12D (3) / G12R (1) / G12S (1) / G12V (4) / G13C (2) / G13D (4) / Q61H (2) / Q61K (1)
	Activating mutation of NRAS[378]	5.9%	Q61K (3) / Q61L (1) / Q61R (1)
	Activating mutation of BRAF [376]	7.1%	G466V(1) / G469A(3) / L597V(1) / V600E(1)
	Activating mutation PIK3CA [379, 380]	4.7%	E542K (1) / E545K (2) / H1047R(1)
	Inactivating mutation PTEN[339]	4.7%	H61R(1) / G251C(1) / R233*(2)

,

Table 5-7 Cancer related and drug related specific mutations in 85 NSCLC cell-lines.

Cell-lines	Disease	Mutated Gene[389, 390]	Type of Mutation	Mutation Details	
				Amino Acid	Nucleotide
A427	NSCLC	KRAS	Activating mutation	G12D	35G>A
A549	NSCLC	KRAS	Activating mutation	G12S	34G>A
Calu1	NSCLC	KRAS	Activating mutation	G12C	34G>T
Calu3	NSCLC	ND			
Calu6	NSCLC	KRAS	Activating mutation	Q61K	181C>A
Colo699	NSCLC	ND *			
DV90	NSCLC	KRAS	Activating mutation	G13D	38G>A
EKVX	NSCLC	ND			
H1155	NSCLC	KRAS	Activating mutation	Q61H	183A>T
H1155	NSCLC	PTEN	Inactivating mutation	R233*	697C>T
H1299	NSCLC	NRAS	Activating mutation	Q61K	181C>A
H1355	NSCLC	KRAS	Activating mutation	G13C	37G>T
H1355	NSCLC	BRAF	Activating mutation	G469A	1406G>C
H1395	NSCLC	BRAF	Activating mutation	G469A	1406G>C
H1437	NSCLC	ND			
H1563	NSCLC	PIK3CA*	Activating mutation	E542K	1624G>A
H1568	NSCLC	ND			
H157	NSCLC	KRAS	Activating mutation	G12R	34G>C
H157	NSCLC	PTEN	Inactivating mutation	G251C	751G>T
H157	NSCLC	PTEN	Inactivating mutation	H61R	182A>G
H1648	NSCLC	ND			
H1650	NSCLC	EGFR	EGFR sensitizing mutation	E746_A750del	2235_2249del15
H1666	NSCLC	BRAF	Activating mutation	G466V	1397G>T

H1734	NSCLC	KRAS	Activating mutation	G13C	37G>T
H1755	NSCLC	BRAF	Activating mutation	G469A	1406G>C
H1770	NSCLC	ND			
H1781	NSCLC	ERBB2*	gefitinib and erlotinib resistant mutation	G776VC	
H1792	NSCLC	KRAS	Activating mutation	G12C	34G>T
H1819	NSCLC	ND			
H1838	NSCLC	ND			
H1915	NSCLC	ND*			
H1944	NSCLC	KRAS*	Activating mutation	G13D	38G>A
H1975	NSCLC	EGFR	EGFR-I sensitizing mutation	L858R	2573T>G
H1975	NSCLC	EGFR	EGFR-I resistant mutation	T790M	2369C>T
H1993	NSCLC	ND			
H2009	NSCLC	KRAS	Activating mutation	G12A	35G>C
H2030	NSCLC	KRAS	Activating mutation	G12C	34G>T
H2052	NSCLC	ND			
H2077	NSCLC	ND*			
H2087	NSCLC	BRAF	Activating mutation	L597V	1789C>G
H2087	NSCLC	NRAS	Activating mutation	Q61K	181C>A
H2110	NSCLC	ND			
H2122	NSCLC	KRAS	Activating mutation	G12C	34G>T
H2126	NSCLC	ND			
H2172	NSCLC	ND*			
H2228	NSCLC	ND			
H23	NSCLC	KRAS	Activating mutation	G12C	34G>T
H23	NSCLC	PTEN	Inactivating mutation	R233*	697C>T
H2347	NSCLC	NRAS	Activating mutation	Q61R	182A>G

H2444	NSCLC	KRAS*	Activating mutation	G12V	
H28	NSCLC	ND			
H2882	NSCLC	ND			
H2887	NSCLC	KRAS*	Activating mutation	G12V	
H3122	NSCLC	ND			
H322	NSCLC	ND			
H3255	NSCLC	EGFR	EGFR-I sensitizing mutation	L858R	34G>T
H358	NSCLC	KRAS	Activating mutation	G12C	34G>T
H441	NSCLC	KRAS	Activating mutation	G12V	35G>T
H460	NSCLC	PIK3CA	Activating mutation	E545K	1633G>A
H460	NSCLC	KRAS	Activating mutation	Q61H	183A>T
H520	NSCLC	ND			
H522	NSCLC	ND			
H596	NSCLC	PIK3CA	Activating mutation	E545K	1633G>A
H647	NSCLC	KRAS	Activating mutation	G13D	38G>A
H661	NSCLC	ND			
H820	NSCLC	EGFR*	EGFR-I sensitizing mutation	E746_T751del, I ins	
H820	NSCLC	EGFR*	EGFR-I resistant mutation	T790M	2369C>T
HCC1171	NSCLC	KRAS*	Activating mutation	G12C	
HCC1195	NSCLC	NRAS*	Activating mutation	Q61L	
HCC1359	NSCLC	ND*			
HCC15	NSCLC	NRAS*	Activating mutation	Q61K	
HCC1833	NSCLC	ND*			
HCC193	NSCLC	ND*			
HCC2279	NSCLC	EGFR*	EGFR-I sensitizing mutation	E746_A750del	2235_2249del15
HCC2429	NSCLC	ND*			

HCC2450	NSCLC	PIK3CK*	Activating mutation	H1047R	3140A>G
HCC2935	NSCLC	EGFR*	EGFR-I sensitizing mutation	E746_A750del, T751A	
HCC364	NSCLC	BRAF	Activating mutation	V600E	1799T>A
HCC366	NSCLC	ND*			
HCC4006	NSCLC	EGFR*	EGFR-I sensitizing mutation	L747_E749del, A750P	
HCC44	NSCLC	KRAS*	Activating mutation	G12C	
HCC461	NSCLC	KRAS*	Activating mutation	G12D	
HCC515	NSCLC	KRAS*	Activating mutation	G13D	
HCC78	NSCLC	ND*			
HCC827	NSCLC	EGFR*	EGFR-I sensitizing mutation	E746_A750del	2235_2249del15
HCC95	NSCLC	ND*			
HOP62	NSCLC	KRAS	Activating mutation	G12C	34G>T
HOP92	NSCLC	ND			
LCLC103H	NSCLC	ND			
LCLC97TM	NSCLC	KRAS	Activating mutation	G12V	35G>T
LouNH91	NSCLC	EGFR*	EGFR-I sensitizing mutation	L747_S752del, P753S	
PC9	NSCLC	EGFR*	EGFR-I sensitizing mutation	E746_A750del	2235_2249del15
SKLU1	NSCLC	KRAS*	Activating mutation	G12D	35G>A

* Mutation was only reported in Ref [390]; # PIK3CA mutation of JIMT-1 was reported by Ref [401]

Abbreviations: ND, no sensitizing/resistant/activating mutation was detected according to COSMIC database and Ref 4

In amplification-based method A1, a NSCLC cell-line is predicted as sensitive to a drug if EGFR in the respective cell-line is amplified [142, 402] and inhibited by the drug at IC50 ≤ 500nM [384]. A gene in a cell-line is considered amplified if its copy number is ≥ 3 [403]. Copy numbers of the evaluated genes in the studied cell-lines were from literatures [403].

In expression-based method E1, a NSCLC cell-line is predicted as sensitive to a drug if EGFR in the respective cell-line is over-expressed [269] and inhibited by the drug at IC50≤500nM [384]. The expression-based method E2 differs from method E1 by an additional condition: all un-inhibited bypass genes in a cell-line are not over-expressed. Bypass genes are frequently up-regulated or amplified in resistant patients [285, 305, 359], which likely enable the promotion of drug-resistant signaling at significant levels. A gene in cancer cell-lines was assumed to be over-expressed if its microarray gene expression level is ≥ 2-fold higher than the lowest level of the same gene in the corresponding healthy tissue cell-lines [404].

5.2.4 Collection of the mutation, ammplification and expression data of NSCLC patients.

37-753 NSCLC patient samples with mutation and amplification data [361, 405, 406] and 45 NSCLC patient samples with expression data [407] were collected from the literatures and Gene Expression Omnibus (GEO) database. Specifically, EGFR somatic mutation data were from the reported study of 58 lung cancer patients from Japan and 61 lung cancer patients from the US [270]. KRAS mutation data were from a reported study of 753 NSCLC patients [361] in which KRAS mutations have been identified by either allele-specific realtime PCR or codon amplification followed by direct sequencing. The

amplification data were from a reported study of 37 NSCLC patients responded clinically to either gefitinib or erlotinib and undergone repeat biopsy and comparative molecular analysis [405]. The gene expression data of 45 NSCLC patients were from the GEO database entry GSE18842 with 45 tumor samples and their paired normal tissue samples analyzed by using the Human Genome U133 Plus 2.0 chip from Affymetrix [407].

5.2.5 Feature selection method

All cell-line data contain the measurements for 22,215 gene probes, The data were subject to the standard preprocessing procedure [408]. The dataset was randomly divided into a training set and an associated testing set of roughly 4:1 ratio. By using repeated random sampling [298], 10,000 training-testing sets, each containing a unique combination of samples (sensitive samples in some combinations are not unique because of the few number of samples), were generated. These 10,000 training-testing sets were randomly placed into 20 sampling groups; each group contains 500 training-testing sets. Every sampling group was then used to derive a set of biomarkers based on consensus scoring and evaluation of gene-ranking consistency of the corresponding 500 training and 500 testing sets. The 20 different signatures derived from these sampling groups were compared to test the level of stability of selected biomarker genes.

SVM, a supervised machine learning method, was used for training a class differentiation system [300, 408, 409]. In classification of microarray datasets, it has been found that supervised machine learning methods generally yield better results [410], particularly for smaller sample sizes [300], and SVM consistently

shows outstanding performance, is less penalized by sample redundancy, and has lower risk for over-fitting [409, 411]. Biomarker genes of each testing-set were selected by using SVM recursive feature elimination (RFE-SVM), which is a wrapper method that selects biomarkers by eliminating non-contributing genes according to a gene-ranking function [119]. Wrapper methods generally perform better than other feature-selection methods [119]. RFE-SVM is the best performing wrapper method and has thus been more widely used in cancer microarray analysis [408, 411].

To further reduce the chance of erroneous elimination of biomarkers due to noises in microarray data, additional gene-ranking consistency evaluation steps were implemented on top of the normal RFE procedures. In step 1, for every testing-set, subsets of genes ranked in the bottom 10% (if no gene was selected in current iteration, this percentage was gradually increased to the bottom 40%) with combined score lower than the first few top-ranked genes were selected such that collective contribution of these genes was less likely outweigh higher-ranked ones. In step 2, for every testing-set, the step-1 selected genes was further evaluated to choose those not ranked in the upper 50% in previous iteration so as to ensure that these genes are consistently ranked lower. In step 3, a consensus scoring scheme was applied to step-2 selected genes such that only those appearing in >90% (if no gene was selected in current iteration, this percentage was gradually reduced to 60%) of the 500 testing-sets were eliminated.

For each sampling-set, different SVM parameters were scanned, various RFE

iteration steps were evaluated to identify the globally optimal SVM parameters and RFE iteration steps that give the highest average class-differentiation accuracy for the 500 testing-sets. The 20 different signatures derived from these sampling-sets were then compared to test the level of stability of selected predictor-genes.

5.3 Result and Discussion

5.3.1 EGFR tyrosine kinase inhibitor bypass signaling in EGFR pathway

The map of EGFR pathway and downstream effectors relevant to cancers is shown in **Figure 5-1**. Overall, 16 bypass signaling routes with 11 bypass genes together with 7 additional bypass signaling routes mediated by the genetic variations of EGFR and 4

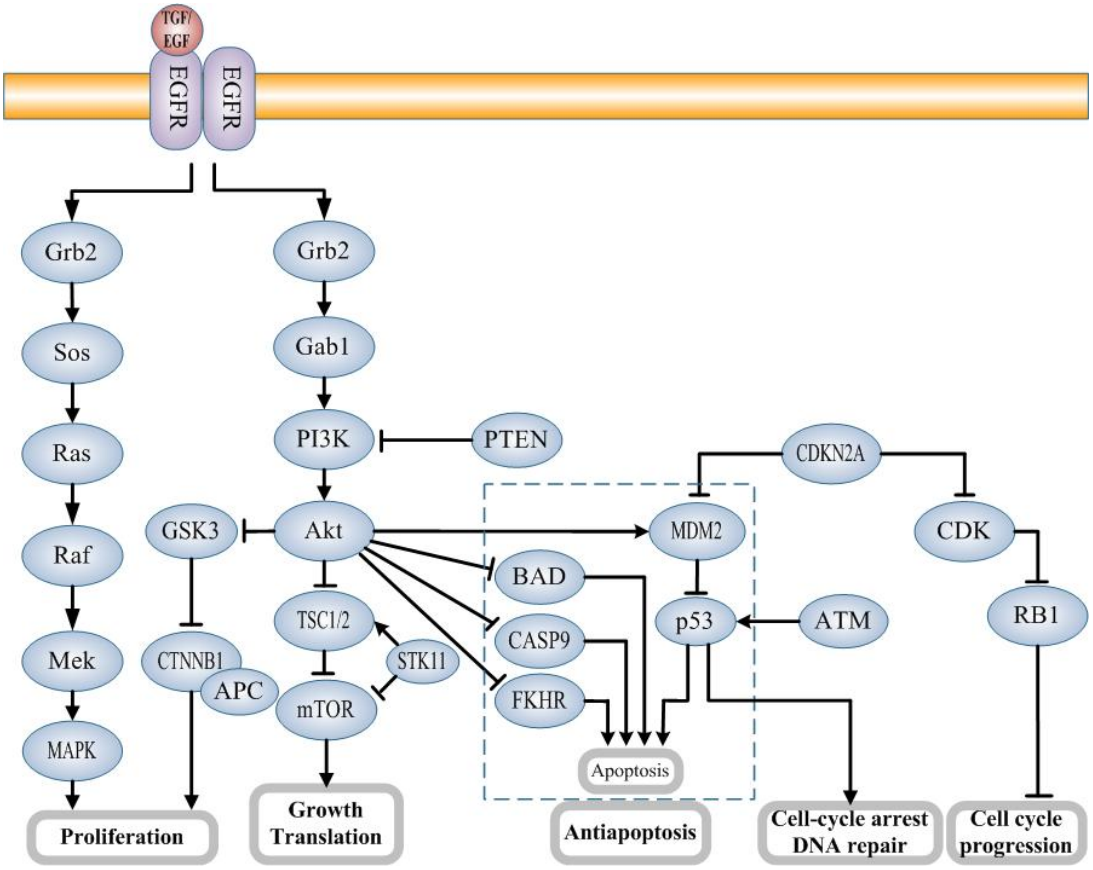


Figure 5-1 The major signaling pathways of the EGFR and downstream effectors relevant to cancers. Modified after Yarden and Sliwkowsk et al (2001),[372] Hynes and Lane (2005),[373] Citri and Yarden (2006),[341] and Normanno et al (2006).[374] Binding of specific ligands (e.g. EGF, heparin-binding EGF, TGF-α) may generate homodimeric complexes resulting in conformational changes in the intracellular EGFR kinase domain, which lead to autophosphorylation and activation. Consequently, signaling molecules, including growth factor receptor-bound protein-2 (Grb-2), Shc and IRS-1 are recruited to the plasma membrane. Activation of several signaling cascades is triggered predominately by the RAS-to-MAPK and the PI3K/Akt pathways, resulting in enhanced tumour growth, survival, invasion and metastasis. Certain mutations in the tyrosine kinase domain may render EGFR constitutively active without their ligands. For cancers with these EGFR activating mutations, the EGFR ligands EGF or TGF-α is unimportant.

downstream drug resistance genes and by enhanced accumulation of internalized EGFR were identified based on the reported experimental evidences that link the specific mutation, amplification, over-expression, or deficiency of each of the 11 bypass-genes, 4 downstream drug resistance genes and EGFR to the resistance of an EGFR tyrosine kinase inhibitor [305, 382]. These are summarized in **Table 5-1** and **5-2**. The bypass signaling events in response to EGFR-I can be divided into three classes. The first class, shown in **Figure 5-2**, involves downstream EGFR-independent signaling via genetic variations of EGFR and downstream drug resistant genes, which have been explored as EGFR-I biomarkers [279-284]. The specific downstream EGFR-independent signaling mechanisms include EGFR mutations resistant to an EGFR-I (D1) [271, 376, 377], activating mutations in Raf (D2), Ras (D3), PI3K (D5), and AKT (D6),[376, 378-380] PTEN loss of function (D4) [339], and enhanced accumulation of internalized EGFR by MDGI (D7) [340]. In **Figure 5-2**, proteins known to carry drug resistant mutations or activating mutations are in darker color and red label. PTEN loss of function is represented by dashed elliptic plate with blank background. Drug resistance mutations in EGFR cause resistance to EGFR-I by such mechanisms as steric impediment of EGFR-I and the increase of ATP affinity to enhance its competitiveness against competitive EGFR-I [377, 412]. While loss-of-function of PTEN can be induced by both PTEN loss[277] and PTEN inactivating mutations [276], PTEN loss occurs significantly more frequently than PTEN inactivating mutations in NSCLC patients [413]. Moreover, PTEN loss alone is not a sufficiently good biomarker for NSCLC [413], indicating the need for collective consideration of multiple bypass mechanisms in predicting drug response.

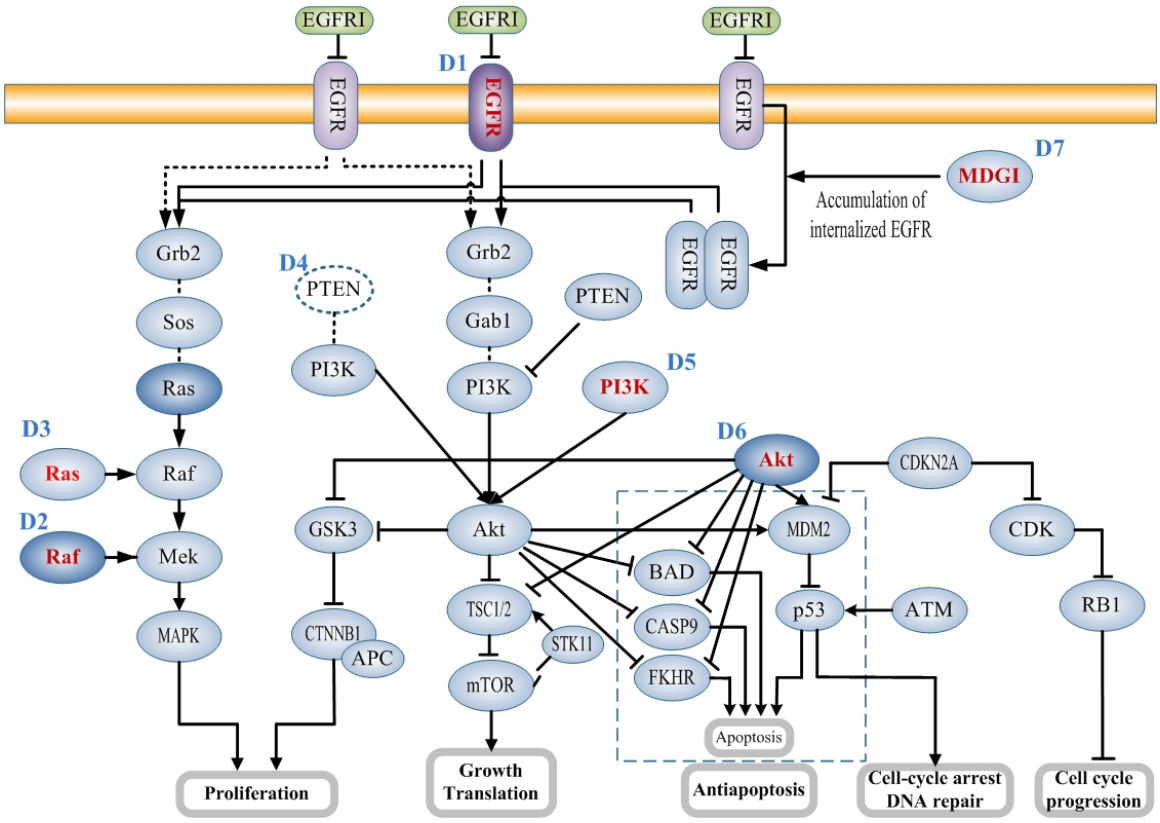


Figure 5-2 EGFR pathway shows EGFR tyrosine kinase inhibitor (EGFRI) bypass mechanisms duo to downstream EGFR-independent signaling involving mutations resistant to EGFRI (D1), activating mutations in Raf (D2), Ras (D3), PI3K (D5), and AkT (D6), PTEN loss of function (D4), and enhanced accumulation of internalized EGFR by MDGI (D7). Proteins known to carry drug resistant mutations or activating mutations are in darker color and red label. The loss of function of PTEN is represented by dashed elliptic plate.

The second class, shown in **Figure 5-3**, involves compensatory signaling due to EGFR transactivation by HER2 (C1) [341, 342], MET (C2) [286, 347], IGF1R (C3) [345], Integrin β1 (C4) [356], and HER3 (C4) [285]. In particular, C3, C4 and C5 activate PI3K via IRS1/IRS2, FAK or a PP2-sensitive kinase, and direct interaction respectively, which are different from the Grb2-Gab1-PI3K path used by the canonical EGFR signaling pathway. The ligands of EGFR and some other receptor tyrosine kinases, particularly members of ErbB families, have the ability to induce not only their own receptor homodimers but also heterodimers with other

selected receptor tyrosine kinases and GPCRs [414, 415], which significantly expands the signaling potential of EGFR [414, 415] and enables the bypass of EGFR inhibition in tumors [285, 286, 341, 342, 345, 347]. Depending on the activating ligand, EGFR and other selected receptor tyrosine kinases are able to form various homodimers or heterodimers with different biological signaling capacities and with response to the inhibition of EGFR or downstream proteins.

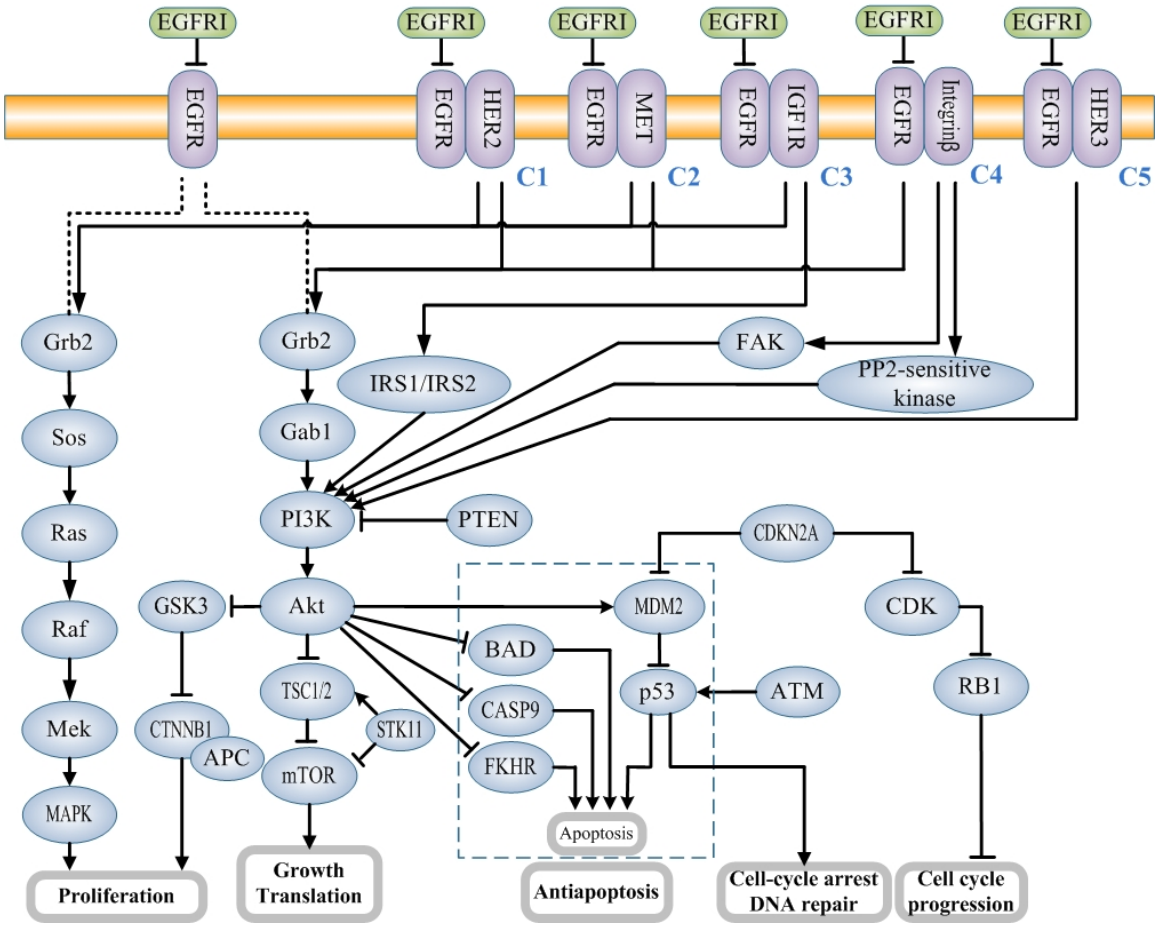


Figure 5-3 EGFR pathway shows EGFR tyrosine kinase inhibitor (EGFRI) bypass mechanisms duo to compensatory signaling of EGFR transactivation with HER2 (C1), MET (C2), IGF1R (C3), Integrinβ1 (C4), and HER3 (C5). In particular, C3, C4 and C5 activates PI3K via IRS1/IRS2, FAK or a PP2-sensitive kinase, and direct interaction respectively

The third class, shown in **Figure 5-4**, involves alternative signaling of VEGFR2 activation (A1) [352], HER2-MET transactivation (A2) [286, 347], PDGFR activation (A3) [350], IGF1R activation (A4) [346], HER2-HER3 transactivation (A5) [342, 343], HER2-HER4 transactivation (A6) [342, 343], MET-HER3 transactivation (A7) [305], PDGFR-HER3 transactivation (A8) [286], Integrin α/β activation (A9) [356], IL6 activation of IL6R-GP130 complex (A10) [358], and Cox2 mediated activation of EP receptors (A11) [360]. In particular, VEGFR activates Raf and Mek via PLC-PKC path and activates PI3K via Shb-FAK path, IGFR activates PI3K via IRS1/IRS2, and HER2-HER3, HER2-HER4, MET-HER3, and PDGFR-HER3 hetrodimers activate PI3K directly, which are different from the Grb2-Sos-Ras_Raf-Mek and Grb2-Gab1-PI3K paths used by the canonical EGFR signaling pathway. The paths A9, A10, and A11 are via non-kinase receptors with a certain downstream protein activating MEK and/or AKT pathways.

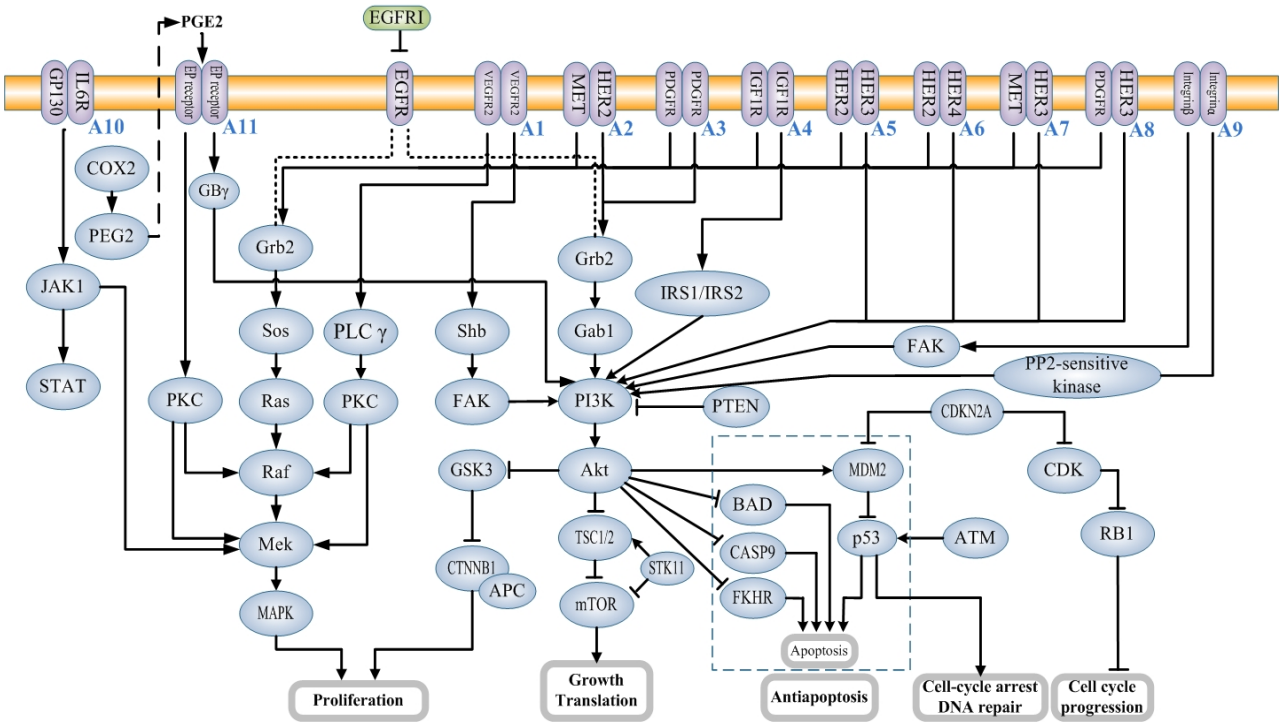


Figure 5-4 EGFR pathway shows EGFR tyrosine kinase inhibitor (EGFR-I) bypass mechanisms duo to alternative signaling of VEGFR2 activation (A1), HER2-MET transactivation (A2), PDGFR activation (A3), IGF1R activation (A4), HER2-HER3 transactivation (A5), HER2-HER4 transactivation (A6), MET-HER3 transactivation (A7), PDGFR-HER3 transactivation (A8), Integrin β1 activation (A9), IL6 activation of IL6R-GP130 complex (A10), and Cox2 mediated activation of EP receptors (A11). In particular, VEGFR activates Raf and Mek via PLCγ-PKC path and activates PI3K via Shb-FAK path, IGFR activates PI3K via IRS1/IRS2, and HER2-HER3, HER2-HER4, MET-HER3, and PDGFR-HER3 hetrodimers activate PI3K directly. The paths A9, A10, and A11 are via non-kinase receptors.

5.3.2 Drug response prediction by genetic and expression profiling of NSCLC cell-lines

The performance and clinical relevance of the individual profile methods M1, M2, M3, A1, E1 and E2 and combination methods CB1 to CB7 in predicting gefitinib, erlotinib and sensitive and resistant NSCLC cell-lines were evaluated. The results are summarized in **Table 5-8**, and are detailed in **Table 5-9** together with the respective cell-line sensitivity data. The methods M2 and M3 correctly predicted 87.5%~100% EGFR-I sensitive and 52.2%~ 65.8% of EGFR-I resistant cell-lines,

which are comparable to the reported 77%~82% accuracies in predicting EGFR-I sensitive, and 54%~87% accuracies in predicting EGFR-I resistant patients [269, 362]. The method A1 correctly predicted 50%~80% EGFR-I sensitive and 62.2%~84.2% of EGFR-I resistant cell-lines respectively, which are comparable to the reported 61% and 74% accuracy in predicting EGFR-I sensitive and resistant patients by the EGFR amplification method [416]. Thus, some of the evaluated single-profile methods are capable of predicting EGFR-I sensitivity from NSCLC cell-lines at performance levels that reflect the sensitivity of real patients.

Both the reported studies and our analyses in Table 1 indicated that the individual-profile tends to show good performance for sensitive cell-lines at the expense of resistant cell-lines or vice versa. Combined mutation and amplification profiles have shown good correlation with clinical response [417]. It is of interest to evaluate whether more balanced performance can be achieved by using

Table 5-8 Ratio of gefitinib, erlotinib, or lapatinib sensitive and resistant NSCLC cell-lines correctly predicted by mutation-based method M1, M2, and M3, amplification-based method A1, expression-based method E1 and E2, and combination methods CB1, CB2, CB3, CB4, CB5, CB6, and CB7. R_S and R_R is ratio of correctly predicted sensitive and resistant cell-lines respectively.

Drug (Efficacy Targets)	Number of Cell-lines (Sensitive/ Resistant)		Ratio of Correctly Predicted Sensitive Cell-Lines (R_S) and Resistant Cell-Lines (R_R) by Different Methods												
			Mutation-Based Method			Amplification-Based Method	Expression-Based Method		Combination of Two Methods					Combination of Three Methods	
			M1	M2	M3	A1	E1	E2	CB1= M3+A1	CB2= M3+E1	CB3= M3+E2	CB4= A1+E1	CB5= A1+E2	CB6= M3+A1+E1	CB7= M3+A1+E2
Gefitinib (EGFR)	44(6/38)	R_S	6/6	6/6	6/6	4/6	3/6	2/6	4/6	3/6	2/6	5/6	5/6	5/6	4/6
		R_R	2/38	23/38	25/38	32 /38	36/38	37/38	32/38	36/38	37/38	31/38	35 /38	31/38	35/38
Erlotinib (EGFR)	51(5/46)	R_S	5/5	5/5	5/5	4/5	3/5	2/5	4/5	3/5	2/5	5/5	5/5	5/5	4/5
		R_R	2 /46	24/46	27/46	38 /46	44 /46	45 /46	40 /46	43/46	45/46	38/46	43/46	39/46	43/46
Lapatinib (HER2, EGFR)	48(8/40)	R_S	8/8	8/8	7/8	4/8	3/8	2/8	4/8	3/8	2/8	5/8	4/8	5/8	3/8
		R_R	2/40	21/40	22 /40	31/40	38 /40	39 /40	33/40	38/40	39 /40	31/40	36/40	31 /40	36/40

combination-profile methods. We evaluated 5 two-profile methods: M3+A1 (CB1), M3+E1 (CB2), M3+E2 (CB3), A1+E1 (CB4), and A1+E2 (CB5), and 2 three-profile methods: M3+A1+E1 (CB6) and M3+A1+E2 (CB7). M3 was used in

these combination methods because it covers the most of the known mutation-based EGFR-I resistance mechanisms. The results are summarized in **Table 5-8**, and provided in **Table 5-9** which also include the cell-line sensitivity data and the genetic and expression profiles of the main target, bypass genes and downstream signaling and regulatory genes. Over-expression of the bypass gene HER2 in NSCLC cell-lines is not expected to significantly contribute to lapatinib resistance because the drug inhibits EGFR and HER2.

Overall, the two-profile combination method CB5 and the three-profile combination method CB7 showed more balanced and improved predictive performance to EGFR-I gefitinib, erlotinib, and lapatinib over the individual-profile methods. Collective consideration of EGFR amplification or over-expression together with the profiles of downstream drug-resistant genes and bypass signaling genes substantially improved the predictive performance for sensitive cell-lines. The performance of multiple profile methods may be affected by at least two factors. One is the substantial level of redundancy among drug sensitizing mutation, amplification and expression profiles and among drug resistant activating/inactivating mutation and expression profiles. Another is the high noise levels of microarray gene expression data[300] that may in some cases negatively affect the performances of the combination methods with expression profiles.

Table 5-10 shows the distribution and coexistence of drug sensitizing mutation, amplification and expression profiles, and drug resistant mutation and expression profiles in the evaluated NSCLC cell-lines. In NSCLC cell-lines, EGFR-I resistance profiles are dominated by RAS activating mutation and HER3 over-expression, which are consistent with literature reports [269, 285]. Our results

show that a gefitinib-sensitive cell-line H3255 [270, 392] was, as predicted, resistant to all of the studied EGFR-Is due to the over-expression of COX2. H3255 is able to acquire resistance to gefitinib by prolonged exposure of the cell to gefitinib in vitro introducing drug resistant mutation, T790M, at EGFR kinase domain [341]. However, as suggested in **Figure 5-4**, such resistance may also be mediated by COX2 induced activation of EP receptors and continued ErbB-3/PI3K/Akt signaling.

Table 5-9 The genetic and expression profiles of the main target, downstream genes and regulator, and bypass genes of 53 NSCLC cell-lines, and the predicted and actual sensitivity of these cell-lines against 3 kinase inhibitors: gefitinib (D1), erlotinib (D2), and lapatinib (D3).

NSCLC Cell lines	**Profile of Main Target (EGFR) Related to Drug Sensitivity**				**Profile of Main Target (EGFR) Related to Drug Resistance**	**Profile of Downstream Signaling Gene or Regulator Directly Contributing to Drug Resistance**				**Profile of Bypass Gene Directly Contributing to Drug Resistance**							**Predicted (Pre) and Actual (Act) Sensitivity to Gefitinib (D1) and Erlotinib (D2)**			**Predicted (Pre) and Actual (Act) Sensitivity to Lapatinib (D3)**	
	over exp	amp (copy no>4)	amp (copy no>3)	s-mut	r-mut	RAS a-mut	BRAF a-mut	PIK3CA a-mut	PTEN loss	HER2 over exp (Not applicable to D3)	HER3 over exp	FGFR1 over exp	IGF1R over exp	VEGFR2 over exp	c-MET over exp	PDGFR over exp	Pre by M1, M2, M3, A1, E1, E2, C1, C2, C3, C4, C5, C6, C7	Act (D1)	Act (D2)	Pre by M1, M2, M3, A1, E1, E2, C1, C2, C3, C4, C5, C6, C7	Act (D3)
Calu3										1	1						R,S,R,R,R,R,R,R,R,R,R,R,R	S	NA	R,S,R,R,R,R,R,R,R,R,R,R,R	S
H3255	1	1	1	1													S,S,S,S,S,S,S,S,S,S,S,S,S	S	S	S,S,S,S,S,S,S,S,S,S,S,S,S	S
HCC2279		1	1	1													S,S,S,S,R,R,S,S,S,S,S,S,S	S	S	S,S,S,S,R,R,S,S,S,S,S,S,S	R
HCC2935	1			1													S,S,S,R,S,S,S,S,S,S,S,S,S	S	S	S,S,S,R,S,S,S,S,S,S,S,S,S	S
HCC4006		1	1	1													S,S,S,S,R,R,S,S,S,S,S,S,S	S	S	S,S,S,S,R,R,S,S,S,S,S,S,S	S
HCC827	1	1	1	1													S,S,S,S,S,S,S,S,S,S,S,S,S	S	S	S,S,S,S,S,S,S,S,S,S,S,S,S	S
A549						1											R,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,R,R,R,R,R,R,R,R,R,R,R,R	R
Calu1						1											R,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,R,R,R,R,R,R,R,R,R,R,R,R	R
Calu6						1											R,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,R,R,R,R,R,R,R,R,R,R,R,R	R
H1299						1											R,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,R,R,R,R,R,R,R,R,R,R,R,R	R
H1355						1											R,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,R,R,R,R,R,R,R,R,R,R,R,R	R
H1395							1										R,S,R,R,R,R,R,R,R,R,R,R,R	R	R	R,S,R,R,R,R,R,R,R,R,R,R,R	R
H1437											1						R,S,R,R,R,R,R,R,R,R,R,R,R	R	R	R,S,R,R,R,R,R,R,R,R,R,R,R	R

H157						1										R,R,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,R,R,R,R,R,R,R,R,R,R,R,R,R	R
H1648															1	R,S,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,S,R,R,R,R,R,R,R,R,R,R,R,R	S
H1650			1	1												S,S,S,R,R,R,S,S,R,R,R,S,S	R	R	S,S,S,R,R,R,S,S,R,R,R,S,S	R
H1666							1				1					R,S,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,S,R,R,R,R,R,R,R,R,R,R,R,R	S
H1770																R,S,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,S,R,R,R,R,R,R,R,R,R,R,R,R	
H1792			1			1									1	R,R,R,R,R,R,R,R,R,R,R,R,R,R	NA	R	R,R,R,R,R,R,R,R,R,R,R,R,R,R	R
H1819			1							1	1					R,S,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,S,R,R,R,R,R,R,R,R,R,R,R,R	S
H1975			1	1	1											S,S,R,R,R,R,R,R,R,R,R,R,R,R	R	R	S,S,R,R,R,R,R,R,R,R,R,R,R,R	R
H1993															1	R,S,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,S,R,R,R,R,R,R,R,R,R,R,R,R	R
H2009						1									1	R,R,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,R,R,R,R,R,R,R,R,R,R,R,R,R	R
H2052															1	R,S,R,R,R,R,R,R,R,R,R,R,R,R	NA	R	R,S,R,R,R,R,R,R,R,R,R,R,R,R	R
H2087						1	1				1					R,R,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,R,R,R,R,R,R,R,R,R,R,R,R,R	R
H2122						1					1					R,R,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,R,R,R,R,R,R,R,R,R,R,R,R,R	R
H2126											1					R,S,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,S,R,R,R,R,R,R,R,R,R,R,R,R	R
H23						1										R,R,R,R,R,R,R,R,R,R,R,R,R,R	NA	R	R,R,R,R,R,R,R,R,R,R,R,R,R,R	R
H2347						1					1				1	R,R,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,R,R,R,R,R,R,R,R,R,R,R,R,R	R
H28																R,S,R,R,R,R,R,R,R,R,R,R,R,R	NA	R	R,S,R,R,R,R,R,R,R,R,R,R,R,R	R
H2882																R,S,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,S,R,R,R,R,R,R,R,R,R,R,R,R	R
H2887						1										R,R,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,R,R,R,R,R,R,R,R,R,R,R,R,R	R
H3122											1					R,S,R,R,R,R,R,R,R,R,R,R,R,R	NA	R	R,S,R,R,R,R,R,R,R,R,R,R,R,R	R
H322			1								1					R,S,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,S,R,R,R,R,R,R,R,R,R,R,R,R	R
H358						1										R,R,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,R,R,R,R,R,R,R,R,R,R,R,R,R	R

H441						1					1			1			R,R,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,R,R,R,R,R,R,R,R,R,R,R,R,R	R
H460						1		1									R,R,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,R,R,R,R,R,R,R,R,R,R,R,R,R	R
H661																1	R,S,R,R,R,R,R,R,R,R,R,R,R,R	NA	R	R,S,R,R,R,R,R,R,R,R,R,R,R,R	R
H820			1	1	1						1						S,S,R,R,R,R,R,R,R,R,R,R,R,R	R	R	S,S,R,R,R,R,R,R,R,R,R,R,R,R	R
HCC1171						1					1						R,R,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,R,R,R,R,R,R,R,R,R,R,R,R,R	R
HCC1195			1			1					1						R,R,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,R,R,R,R,R,R,R,R,R,R,R,R,R	R
HCC1359																1	R,S,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,S,R,R,R,R,R,R,R,R,R,R,R,R	R
HCC15						1											R,R,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,R,R,R,R,R,R,R,R,R,R,R,R,R	R
HCC1833											1						R,S,R,R,R,R,R,R,R,R,R,R,R,R	NA	R	R,S,R,R,R,R,R,R,R,R,R,R,R,R	R
HCC193	1		1												1		R,S,R,R,S,R,R,S,R,S,R,S,R	R	R	R,S,R,R,S,R,R,S,R,S,R,S,R	R
HCC2429																	R,S,R,R,R,R,R,R,R,R,R,R,R,R	NA	R	R,S,R,R,R,R,R,R,R,R,R,R,R,R	R
HCC2450								1			1						R,S,R,R,R,R,R,R,R,R,R,R,R,R	NA	R	R,S,R,R,R,R,R,R,R,R,R,R,R,R	R
HCC366	1																R,S,R,R,S,S,R,R,S,S,S,S,S	R	R	R,S,R,R,S,S,R,R,S,S,S,S,S	R
HCC44						1											R,R,R,R,R,R,R,R,S,R,R,R,R,R	R	R	R,R,R,R,R,R,R,R,S,R,R,R,R,R	R
HCC461						1											R,R,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,R,R,R,R,R,R,R,R,R,R,R,R,R	R
HCC515						1					1						R,R,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,R,R,R,R,R,R,R,R,R,R,R,R,R	R
HCC78											1						R,S,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,S,R,R,R,R,R,R,R,R,R,R,R,R	R
HCC95																	R,S,R,R,R,R,R,R,R,R,R,R,R,R	R	R	R,S,R,R,R,R,R,R,R,R,R,R,R,R	R

Notes: “1” indicates the corresponding profile is positive (over-expressed, amplified or mutated) and blank indicates negative (not over-expressed, amplified or mutated) respectively. “S”, “R”, “NA”, “s-mut”, “r-mut”, ‘a-mut’, “amp”, “over exp”, “pre”, and “act” stands for sensitive to drug, resistant to drug, no available drug sensitivity, drug sensitive mutation, drug resistance mutation, activating mutation, amplification, over expression, predicted drug sensitivity, and actual drug sensitivity respectively. The prediction methods M1, M2, M3, A1, E1, E2, C1, C2, C3, C4, C5, C6, and C7 are described in the text.

Table 5-10 The distribution and coexistence of amplification and expression profiles, and the drug resistance mutation and expression profiles in NSCLC cell-lines.

Cancer: NSCLC **Main Target for the Treatment of Specific Cancer:** EGFR **Drugs Evaluated:** gefitinib (D1), erlotinib (D2), and lapatinib (D3)																		
Drug Sensitizing or Resistance Profile (*index*)	**Number of Cell-Lines with This Profile**	**Number of These Cell-Lines with Another Sensitizing Profile**		**Number of These Cell-Lines with Another Resistance Profile**												**Number of These Cell-Lines Sensitive/ Resistant to Drug**		
		Drug Sensitizing Profile		***Drug Resistance Profile***														
Drug Sensitizing profile		***S1***	***S2***	***R1***	***R2***	***R3***	***R4***	***R5***	***R6***	***R7***	***R8***	***R9***	***R10***	***R11***	***R12***	***D1***	***D2***	***D3***
EGFR amp(copy no≥3) (***S1***)	12		3	2	2				1	4	1	1				4/7	4/8	4/8
EGFR over exp (***S2***)	5	3									1					3/2	3/1	3/2
Drug Resistance profile																		
EGFR r-mut (***R1***)	2									1						0/2	0/2	0/2
RAS a-mut (***R2***)	22					1				7	2				1	0/20	0/22	0/21
BRAF a-mut (***R3***)	3				1					1						0/3	0/3	1/2
PIK3CA a-mut (***R4***)	2									2						0/1	0/2	0/2
PTEN loss (***R5***)	0																	
HER2 over exp (***R6***)	2									2						1/1	0/1	2/0
HER3 over exp (***R7***)	18			1	7	1	2		2		1				1	1/14	0/17	3/12
MET over exp (***R8***)	5				2					1						0/5	0/5	1/4
PDGFR over exp (***R9***)	4																	
IGF1R over exp (***R10***)	0																	

FGFR1 over exp (***R11***)	0																
VEGFR2 over exp (***R12***)	1			1					1						0/1	0/1	0/1

5.3.3 Relevance and limitations of cell-line data for drug response studies

The relevance of our studied cell-lines for EGFR-I response studies was evaluated by comparison of the distribution profiles of drug resistant mutations of EGFR and 3 downstream genes and the expression profiles of EGFR and 8 bypass genes in our studied 53 NSCLC cell-lines (including all sensitive and resistant cell-lines against the three drugs) and those of 45 NSCLC patient samples with expression data [407] and 37-753 NSCLC patient samples with mutation and amplification data [361, 405, 406] to determine if the cell-line profiles are sufficiently close to those of real patient samples. As shown in **Table 5-11**, the percentages of our studied NSCLC cell-lines carrying drug-resistant mutant genes and up-regulated bypass genes tend to be lower but roughly comparable to those of the patient samples with the exception of VEGFR2. The level of VEGFR2 elevation in patient samples is significantly higher than that of cell-lines. It has been reported that, in some circumstances, the development of VEGFR2 expressing cells is strongly influenced by VEGFR2 ligands in the microenvironment [418]. The VEGFR2 ligand VEGF has been found to be over-expressed in association with microlocalisation of M1 and M2 macrophages in NSCLC [419], which likely contribute to the over-expression of VEGR2 in larger percentages of NSCLC patients [420-422]. Apart from VEGFR2 ligands, tumor microenvironment in NSCLC is known to promote the expression of a number of other factors such as CD163, HLA-DR, iNOS, MRP 8/14, and hypoxia related proteins [419, 423]. The effects of these

microenvironment-related factors in upregulating VEGFR2 and other bypass genes may not be fully reflected by the cell-line based models, which partly contribute to the discrepancy between the expression profiles of some of the genes in our studied NSCLC cell-lines and those of real NSCLC patient samples.

5.3.4 The usefulness of cell-line expression data for identifying drug response biomarkers

We used our support vector machines recursive feature elimination (SVM-RFE) method [301] to select EGFR-I response biomarkers from the gefitinib, erlotinib, or lapatinib sensitive and resistant NSCLC cell-lines and compare them with the published EGFR-I response biomarkers derived from NSCLC patient samples [371] and cell-lines [293, 328] by using the differentially expressed genes method. In our method, NSCLC cell-lines sensitive and resistant to each EGFR-I were separated by using SVM classification and the gene expression of these cell-lines. The contribution of each gene to the separation of sensitive and resistant cell-lines was ranked by the SVM feature ranking algorithm and the least contributing genes were selected and subsequently eliminated by consensus scoring from repeated random sampling [298] and by incorporating a multi-step gene-ranking consistency evaluation procedure into the SVM-RFE method [301].

Table 5-11 Comparison of the expression profiles of EGFR and bypass genes in NSCLC cell-lines and patient samples.

Target or Bypass gene	% of the 53 NSCLC cell lines with the gene up-regulated or lost	% of the 45 NSCLC patients with the gene up-regulated or lost	Target or Bypass gene	% of the 53 NSCLC cell lines with the gene harboring resistance mutation or amplified	% of the 37-753 NSCLC patients with the gene harboring resistance mutation or amplified
EGFR	Up 9.4%	Up 28.9%	EGFR	Mut 3.8%	Mut 19%-49%
HER2	Up 3.8%	Up 15.6%	PIK3CA	Mut 3.8%	Mut 5%
HER3	Up 33.9%	Up 22.2%	Ras	Mut 41.4%	Mut 20%
IGF1R	Up 0.0%	Up 4.4%	EGFR	Amp 22.6%	Amp 8%
c-MET	Up 9.4%	Up 8.9%			
PDGFR	Up 7.6%	Up 24.4%			
FGFR	Up 0.0%	Up 0.0%			
VEGFR2	Up 1.9%	Up 51.1%			
PTEN	Loss 0.0%	Loss 8.9%			

The statistics of SVM-RFE selected gefitinib, erlotinib, and lapatinib response biomarkers and those of the published studies [293, 328, 371] are summarized in **Table 5-12** and the detailed lists of the selected biomarkers are provided in the **Appendix B Tables 1, 2** and **3** respectively. The numbers of our selected biomarkers (65-148) are comparable to those of the published studies (51-332) [293, 328, 371], but few biomarker genes are commonly selected by these studies. For instance, only 0 and 5 of the 148 gefitinib response biomarkers selected in this work are commonly selected by the published studies from NSCLC patient samples [293] and cell-lines [328] respectively, and none of the 51 biomarkers selected by the published study from NSCLC patient samples [293] is commonly selected by the other published study from NSCLC cell-lines [328]. This discrepancy arises in part from the small number of samples used in each study as well as the intrinsic noise in the gene expression data [298-302]. It is noted that our selected biomarkers contain significantly higher number of drug target and bypass genes (4-5) than

those (0-1) of the published studies [293, 328, 371], which may result from the use of significantly higher number of samples (44-53) in our study than those (11-25) in the published studies [293, 328, 371], and the use of multiple sampling, consensus scoring and multi-step gene-ranking consistency evaluation procedure. The ability in identifying substantial number of drug target and bypass genes as biomarkers provides further evidence about the usefulness of cell-line data in facilitating the discovery of drug response biomarkers [303-306].

ERBB3 has been repeatedly selected as biomarkers for the three evaluated drugs by more than one method. Recent investigations have shown that ERBB3 is responsible for tumor resistance to therapeutic agents targeting EGFR or HER2 [424], and is associated with poor prognosis [425], decreased survival in patients with early stage [426] and overall survival [427] in NSCLC. Therefore, it is not surprising that ERBB3 was repeatedly selected. ERBB3 is kinase inactive and thus is not an easily druggable target for developing small molecule inhibitors [424]. RNA aptamers and siRNAs targeting ERBB3 may be explored as potential therapeutics in combination with EGFR and HER2 targeted drugs [428, 429].

The testing accuracies of the RFE-SVM method in differentiating Gefitinib, Erlotinib and Laptinib resistant/sensitive cell-lines are 76.3%, 87.3% and 71.2% respectively. The reported testing accuracies of the method B and C in differentiating Gefitinib resistant/sensitive patients are 91.7% and 100% [293, 371], and that of the method E in differentiating Erlotinib resistant/sensitive cell-lines is 80% [328]. Our RFE-SVM method appears to perform slightly better for Erlotinib but worse for Gefitinib than those of the existing methods. It is cautioned that it is not appropriate to straightforwardly compare these methods based on our and reported testing results, because these tests are based on separate

testing samples that are very small in sizes and high in size differences (44-53 samples in our study vs 5-12 in the reported studies) with the relevant data containing high genetic and measurement variations. Further tests based on a common set of more diverse samples may enable more appropriate comparison of these methods.

Table 5-12 Statistics of the SVM-RFE selected gefitinib, erlotinib, and lapatinib response biomarkers in comparison with those of the published studies.

Drug	**No of resistant/ sensitive NSCLC cell-lines / patients**	**Method**	**No of biomarkers selected by method**	**No of biomarkers also selected by another method**	**Bypass genes selected as biomarker by method**
Gefitinib	38/6	A: SVM-RFE; This work	148	0 by B; 5 by C	ERBB3, EGFR, FGFR1, MET
	7/10 (patients)	B: Differential expression method[371]	51	0 by A; 0 by C	
	6/5	C: Differential expression method[293]	332	5 by A; 0 by C	ERBB3
Erlotinib	46/7	D: SVM-RFE; This work	65	3 by E	EGFR, ERBB3, FGFR3, MET
	11/14	E: Differential expression method[328]	180	3 by A	EGFR
Lapatinib	40/8	F: SVM-RFE; This work	98		ERBB2, ERBB3, FGFR3, PDGFR, COX2

5.4 Conclusion

Consideration of bypass signaling from pathway regulation perspectives appears to be highly useful for deriving knowledge-based drug response biomarkers to effectively predict drug responses as well as for understanding the mechanism of pathway regulation and drug response. The bypass signaling based biomarkers described in this and other studies can be experimentally validated by the methods used for discovering compensatory PI3K/Akt/mTor activation [304], MET amplification [305], and CRAF overexpression [306] as a mechanism of acquired resistance to imatinib in CML, EGFR-I therapy in NSCLC [305], and BRAF inhibitor therapy in melanomas [306] respectively. Specifically, mutation, copy number or expression analysis is conducted in both resistant and sensitive samples to determine if the biomarkers show resistant mutations, marked amplifications or elevated expressions in resistant samples only, followed by the investigation of whether inhibition or down-regulation of the bypass gene together with the inhibition or down-regulation of the drug target reduces the resistance effects.

The currently available molecular interaction, network signaling and regulation, and the genetic and expression profiles of regulatory genes appear to have reached the level for facilitating the discovery of some of the drug response biomarkers based on various drug resistance mechanisms. The mining of the relevant information from the literatures is a key step towards the identification of bypass signalling based drug response biomarkers, which may be hindered by multiple gene and protein nomenclatures [430] and complexities of languages [431]. Collective use of dictionary-based and vocabulary-based methods, disambiguation and correct classification algorithms, and restricted domain search strategies may

enable more comprehensive and efficient search of the relevant information [431].

The drug sensitivity prediction capability of biomarkers may be affected by multiple factors including the number and quality of samples,[303] the genetic variation [331, 332], plasticity [333], and microenvironment [334, 335] of the samples, and the quality and stability of the predicted biomarkers [432]. In addition to the collection of sufficiently diverse patient samples, collective analysis of mutation, amplification and expression profiles of target, bypass genes, and downstream drug-resistant genes is potentially useful for facilitating drug sensitivity prediction. Development, integration and expanding application of next generation sequencing [433], microarrays [434], and copy number variation [435] detection tools and methods coupled with expanded knowledge of systems biology, cancer biology and drug resistance bypass mechanisms and the further improvement of biomarker discovery methods [301, 432, 436] enable more accurate prediction of drug sensitivity.

Chapter 6 Concluding Remarks

6.1 Major findings and merits

6.1.1 Merits of A two-step Target Binding and Selectivity Support Vector Machines Approach for Virtual Screening of Dopamine Receptor Subtype-Selective Ligands

In this work, we introduced a new two-step support vector machines target-binding and selectivity screening method for searching DR subtype-selective ligands, which was tested together with three previously-used machine learning methods for searching D1, D2, D3 and D4 selective ligands. Its subtype selective ligand identification rates are significantly better than, and its multi-subtype ligand identification rates are comparable to the best rates of the previously used methods. Our method produced low false-hit rates in screening 13.56M PubChem, 168,016 MDDR and 657,736 ChEMBLdb compounds. Molecular features important for subtype selectivity were extracted by using the recursive feature elimination feature selection method. These features are consistent with literature-reported features. Our method showed similar performance in searching estrogen receptor subtype selective ligands.

Virtual screening methods have been increasingly explored for facilitating the discovery of target selective drugs for enhanced therapeutics and reduced side effects. Our study further suggested that the two-step target binding and selectivity support vector machines virtual screening tools developed from protein subtype ligands with unspecified subtype selectivity are capable of identifying protein subtype selective ligands at good yields, subtype selectivity and low false-hit rates

in screening large chemical libraries.

6.1.2 Merits of Building a prediction model for IKK beta inhibitors

SVM shows substantial capability in identifying IKK beta inhibitors at comparable yield and in many cases substantially lower false-hit rate than those of typical VS tools reported in the literatures and evaluated in this work. 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. 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 for facilitating the discovery of IKK beta inhibitors and other active compounds.

6.1.3 Merits of Analysis of bypass signaling in EGFR pathway and profiling of bypass genes for predicting response to anticancer EGFR tyrosine kinase inhibitors

In this work, we searched and analyzed 16 literature-reported EGFR tyrosine kinase inhibitor bypass signaling routes in EGFR pathway, which include 5 compensatory routes of EGFR transactivation by another receptor, and 11 alternative routes activated by another receptor. Their expression profiles together with the mutational, amplification and expression profiles of EGFR and 4 downstream drug-resistant genes were used as new sets of biomarkers for identifying 53 NSCLC cell-lines sensitive or resistant to EGFR tyrosine kinase inhibitors gefitinib, erlotinib, and lapatinib. The collective profiles of all 16 genes

distinguish sensitive and resistant cell-lines are better than those of individual genes and the combined EGFR and downstream drug resistant genes, and their derived cell-line response rates are consistent with the reported clinical response rates of the three drugs. The usefulness of cell-line data for drug response studies was further analyzed by comparing the expression profiles of EGFR and bypass genes in NSCLC cell-lines and patient samples, and by using a machine learning feature selection method for selecting drug response biomarkers. Our study suggested that the profiles of drug bypass signaling are highly useful for improved drug response prediction.

6.2 Limitations and suggestions for future studies

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

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

Drug response biomarkers facilitate the characterization of patient populations and quantitation of the extent to which new drugs reach intended targets. Clinically useful biomarkers are required to inform regulatory and therapeutic decision-making regarding candidate drugs. Further improvement in measurement quality, annotation accuracy and coverage, and signature-selection will enable the derivation of more accurate signatures for facilitating drug response biomarker and target discovery. The currently available platforms for microarray data are different. Therefore if we could synchronize the platform and provide more samples, we could further improve the accuracy of our system and reduce the computational time. The gene ontology information also could be integrated into the system and the selected genes would be given a biological meaning directly. The drug sensitivity prediction capability of biomarkers may be affected by

multiple factors including the number and quality of samples, the genetic variation, plasticity, and microenvironment of the samples, and the quality and stability of the predicted biomarkers. In addition to the collection of sufficiently diverse patient samples, collective analysis of mutation, amplification and expression profiles of target, bypass genes, and downstream drug-resistant genes is potentially useful for facilitating drug sensitivity prediction. Development, integration and expanding application of next generation sequencing, microarrays, and copy number variation detection tools and methods coupled with expanded knowledge of systems biology, cancer biology and drug resistance bypass mechanisms and the further improvement of biomarker discovery methods enable more accurate prediction of drug sensitivity.

7 BIBLIOGRAPHY

1. Ashburn, T.T. and K.B. Thor, Drug repositioning: Identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 2004. **3**(8): p. 673-683.
2. Sollano, J.A., et al., The economics of drug discovery and the ultimate valuation of pharmacotherapies in the marketplace. *Clin Pharmacol Ther*, 2008. **84**(2): p. 263-6.
3. Newman, D.J., Natural products as leads to potential drugs: an old process or the new hope for drug discovery? *J Med Chem*, 2008. **51**(9): p. 2589-99.
4. Ohlstein, E.H., R.R. Ruffolo, Jr., and J.D. Elliott, Drug discovery in the next millennium. *Annu Rev Pharmacol Toxicol*, 2000. **40**: p. 177-91.
5. Drews, J., Drug discovery: a historical perspective. *Science*, 2000. **287**(5460): p. 1960-4.
6. Friedberg, I., T. Kaplan, and H. Margalit, Evaluation of PSI-BLAST alignment accuracy in comparison to structural alignments. *Protein Sci*, 2000. **9**(11): p. 2278-84.
7. Muller, A., R.M. MacCallum, and M.J. Sternberg, Benchmarking PSI-BLAST in genome annotation. *J Mol Biol*, 1999. **293**(5): p. 1257-71.
8. Chen, C., et al., Predicting protein structural class based on multi-features fusion. *J Theor Biol*, 2008. **253**(2): p. 388-92.
9. Li, Z.R., et al., PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res*, 2006. **34**(Web Server issue): p. W32-7.
10. Cerami, E.G., et al., cPath: open source software for collecting, storing, and querying biological pathways. *Bmc Bioinformatics*, 2006. **7**.
11. Cases, I., et al., CARGO: a web portal to integrate customized biological information. *Nucleic Acids Res*, 2007. **35**(Web Server issue): p. W16-20.
12. Nakazato, T., et al., BioCompass: a novel functional inference tool that utilizes MeSH hierarchy to analyze groups of genes. *In Silico Biol*, 2008. **8**(1): p. 53-61.
13. Worldwide value of bioinformatics; Available from: http://www.bccresearch.com/report/BIO051A.html.
14. Southan, C., P. Varkonyi, and S. Muresan, Complementarity between public and commercial databases: new opportunities in medicinal chemistry informatics. *Curr Top Med Chem*, 2007. **7**(15): p. 1502-8.
15. Rester, U., From virtuality to reality - Virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective. *Curr Opin Drug Discov Devel*, 2008. **11**(4): p. 559-68.
16. Rollinger, J.M., H. Stuppner, and T. Langer, Virtual screening for the discovery of bioactive natural products. *Prog Drug Res*, 2008. **65**: p. 211, 213-49.
17. Shoichet, B.K., Virtual screening of chemical libraries. *Nature*, 2004. **432**(7019): p. 862-5.
18. Lengauer, T., et al., Novel technologies for virtual screening. *Drug Discov Today*, 2004. **9**(1): p. 27-34.
19. Davies, J.W., M. Glick, and J.L. Jenkins, Streamlining lead discovery by aligning in silico and high-throughput screening. *Curr Opin Chem Biol*, 2006. **10**(4): p. 343-51.
20. Willett, P., Similarity-based virtual screening using 2D fingerprints. *Drug Discov Today*, 2006. **11**(23-24): p. 1046-53.
21. van de Waterbeemd, H. and E. Gifford, ADMET in silico modelling: towards prediction paradise? *Nat Rev Drug Discov*, 2003. **2**(3): p. 192-204.
22. Matthew W. B. Trotter , S.B.H., Support Vector Machines for ADME Property Classification. *QSAR & Combinatorial Science*, 2003. **22**(5): p. 533-548.
23. Cavasotto, C.N. and A.J. Orry, Ligand docking and structure-based virtual screening in drug discovery. *Curr Top Med Chem*, 2007. **7**(10): p. 1006-14.
24. Guido, R.V., G. Oliva, and A.D. Andricopulo, Virtual screening and its integration with modern drug design technologies. *Curr Med Chem*, 2008. **15**(1): p. 37-46.
25. Brooijmans, N. and I.D. Kuntz, Molecular recognition and docking algorithms. *Annu Rev Biophys Biomol Struct*, 2003. **32**: p. 335-73.
26. Halperin, I., et al., Principles of docking: An overview of search algorithms and a guide to

scoring functions. *Proteins*, 2002. **47**(4): p. 409-43.
27. Wang, R., Y. Lu, and S. Wang, Comparative evaluation of 11 scoring functions for molecular docking. *J Med Chem*, 2003. **46**(12): p. 2287-303.
28. Moitessier, N., et al., Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br J Pharmacol*, 2008. **153 Suppl 1**: p. S7-26.
29. Warren, G.L., et al., A critical assessment of docking programs and scoring functions. *J Med Chem*, 2006. **49**(20): p. 5912-31.
30. Schulz-Gasch, T. and M. Stahl, Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J Mol Model*, 2003. **9**(1): p. 47-57.
31. Kim, R. and J. Skolnick, Assessment of programs for ligand binding affinity prediction. *J Comput Chem*, 2008. **29**(8): p. 1316-31.
32. Kirchmair, J., et al., 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): p. 213-28.
33. Sheridan, R.P., G.B. McGaughey, and W.D. Cornell, Multiple protein structures and multiple ligands: effects on the apparent goodness of virtual screening results. *J Comput Aided Mol Des*, 2008. **22**(3-4): p. 257-65.
34. Jain, A.N., Bias, reporting, and sharing: computational evaluations of docking methods. *J Comput Aided Mol Des*, 2008. **22**(3-4): p. 201-12.
35. Hawkins, P.C., A.G. Skillman, and A. Nicholls, Comparison of shape-matching and docking as virtual screening tools. *J Med Chem*, 2007. **50**(1): p. 74-82.
36. Wolber, G., et al., Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discov Today*, 2008. **13**(1-2): p. 23-9.
37. Sun, H., Pharmacophore-based virtual screening. *Curr Med Chem*, 2008. **15**(10): p. 1018-24.
38. Carosati, E., et al., Discovery of novel and cardioselective diltiazem-like calcium channel blockers via virtual screening. *J Med Chem*, 2008. **51**(18): p. 5552-65.
39. Moffat, K., et al., A comparison of field-based similarity searching methods: CatShape, FBSS, and ROCS. *J Chem Inf Model*, 2008. **48**(4): p. 719-29.
40. McGaughey, G.B., et al., Comparison of topological, shape, and docking methods in virtual screening. *J Chem Inf Model*, 2007. **47**(4): p. 1504-19.
41. Stahura, F.L. and J. Bajorath, New methodologies for ligand-based virtual screening. *Curr Pharm Des*, 2005. **11**(9): p. 1189-202.
42. Ecker, G.F., T. Stockner, and P. Chiba, Computational models for prediction of interactions with ABC-transporters. *Drug Discov Today*, 2008. **13**(7-8): p. 311-7.
43. Cruz-Monteagudo, M., M.N. Cordeiro, and F. Borges, Computational chemistry approach for the early detection of drug-induced idiosyncratic liver toxicity. *J Comput Chem*, 2008. **29**(4): p. 533-49.
44. Ma, X.H., et al., Advances in machine learning prediction of toxicological properties and adverse drug reactions of pharmaceutical agents. *Curr Drug Saf*, 2008. **3**(2): p. 100-14.
45. Huang, J., et al., Identifying P-glycoprotein substrates using a support vector machine optimized by a particle swarm. *J Chem Inf Model*, 2007. **47**(4): p. 1638-47.
46. Xue, Y., et al., Prediction of P-glycoprotein substrates by a support vector machine approach. *J Chem Inf Comput Sci*, 2004. **44**(4): p. 1497-505.
47. Matter, H., et al., Computational approaches towards the rational design of drug-like compound libraries. *Comb Chem High Throughput Screen*, 2001. **4**(6): p. 453-75.
48. Walters, W.P. and M.A. Murcko, Prediction of 'drug-likeness'. *Adv Drug Deliv Rev*, 2002. **54**(3): p. 255-71.
49. Zernov, V.V., et al., Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J Chem Inf Comput Sci*, 2003. **43**(6): p. 2048-56.
50. Burbidge, R., et al., Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers and Chemistry*, 2001. **26**(1): p. 5-14.
51. Manallack, D.T. and D.J. Livingstone, Neural networks in drug discovery: have they lived up to their promise? *European Journal of Medicinal Chemistry*, 1999. **34**(3): p. 195-208.
52. Trotter, M.W.B. and S.B. Holden, Support vector machines for ADME property classification. *QSAR & Combinatorial Science*, 2003. **22**(5): p. 533-548.
53. Oprea, T.I. and H. Matter, Integrating virtual screening in lead discovery. *Curr Opin Chem Biol*, 2004. **8**(4): p. 349-58.
54. Bocker, A., G. Schneider, and A. Teckentrup, NIPALSTREE: a new hierarchical clustering

approach for large compound libraries and its application to virtual screening. *J Chem Inf Model*, 2006. **46**(6): p. 2220-9.
55. Schuster, D., et al., The discovery of new 11beta-hydroxysteroid dehydrogenase type 1 inhibitors by common feature pharmacophore modeling and virtual screening. *J Med Chem*, 2006. **49**(12): p. 3454-66.
56. Steindl, T., C. Laggner, and T. Langer, Human rhinovirus 3C protease: generation of pharmacophore models for peptidic and nonpeptidic inhibitors and their application in virtual screening. *J Chem Inf Model*, 2005. **45**(3): p. 716-24.
57. Schroeter, T., et al., Machine Learning Models for Lipophilicity and Their Domain of Applicability. *Mol Pharm*, 2007. **4**(4): p. 524-538.
58. Li, H., et al., Machine learning approaches for predicting compounds that interact with therapeutic and ADMET related proteins. *J Pharm Sci*, 2007. **(Published Online)**.
59. Fox, T. and J.M. Kriegl, Machine learning techniques for in silico modeling of drug metabolism. *Curr Top Med Chem*, 2006. **6**(15): p. 1579-91.
60. Duch, W., K. Swaminathan, and J. Meller, Artificial intelligence approaches for rational drug design and discovery. *Curr Pharm Des*, 2007. **13**(14): p. 1497-508.
61. Chen, B., et al., Evaluation of machine-learning methods for ligand-based virtual screening. *J Comput Aided Mol Des*, 2007.
62. Stichtenoth, D.O. and J.C. Frolich, The second generation of COX-2 inhibitors: what advantages do the newest offer? *Drugs*, 2003. **63**(1): p. 33-45.
63. Linkins, L.A. and J.I. Weitz, Pharmacology and clinical potential of direct thrombin inhibitors. *Current Pharmaceutical Design*, 2005. **11**(30): p. 3877-3884.
64. Ribeiro, S. and R. Horuk, The clinical potential of chemokine receptor antagonists. *Pharmacology & Therapeutics*, 2005. **107**(1): p. 44-58.
65. Spaltenstein, A., et al., Discovery of next generation inhibitors of HIV protease. *Current topics in medicinal chemistry*, 2005. **5**(16): p. 1589-1607.
66. Fabbro, D., et al., Protein kinases as targets for anticancer agents: from inhibitors to useful drugs. *Pharmacology & Therapeutics*, 2002. **93**(2-3): p. 79-98.
67. Kumar, R., V.P. Singh, and K.M. Baker, Kinase inhibitors for cardiovascular disease. *Journal of Molecular and Cellular Cardiology*, 2006. **doi:10.1016/j.yjmcc.2006.09.005**.
68. Rotella, D.P., Phosphodiesterase 5 inhibitors: current status and potential applications. *Nature reviews. Drug discovery*, 2002. **1**(9): p. 674-682.
69. Pacher, P. and V. Kecskemeti, Trends in the development of new antidepressants. Is there a light at the end of the tunnel? *Current Medicinal Chemistry*, 2004. **11**(7): p. 925-943.
70. Franke, L., et al., 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): p. 6997-7004.
71. Jorissen, R.N. and M.K. Gilson, Virtual screening of molecular databases using a support vector machine. *J Chem Inf Model*, 2005. **45**(3): p. 549-61.
72. Hert, J., et al., 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): p. 462-70.
73. Chen, B., et al., 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): p. 478-486.
74. Glick, M., et al., 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): p. 193-200.
75. Lepp, Z., T. Kinoshita, and H. Chuman, Screening for new antidepressant leads of multiple activities by support vector machines. *J Chem Inf Model*, 2006. **46**(1): p. 158-67.
76. Harper, G., et al., Prediction of biological activity for high-throughput screening using binary kernel discrimination. *J Chem Inf Comput Sci*, 2001. **41**(5): p. 1295-300.
77. Yamazaki, K., et al., 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): p. 1371-1379.
78. Vidal, D., M. Thormann, and M. Pons, A novel search engine for virtual screening of very large databases. *J Chem Inf Model*, 2006. **46**(2): p. 836-43.
79. Mozziconacci, J.C., et al., Optimization and validation of a docking-scoring protocol; application to virtual screening for COX-2 inhibitors. *J Med Chem*, 2005. **48**(4): p. 1055-68.
80. Vangrevelinghe, E., et al., Discovery of a potent and selective protein kinase CK2 inhibitor

by high-throughput docking. *J Med Chem*, 2003. **46**(13): p. 2656-62.
81. Enyedy, I.J., et al., Discovery of small-molecule inhibitors of Bcl-2 through structure-based computer screening. *J Med Chem*, 2001. **44**(25): p. 4313-24.
82. Doman, T.N., et al., Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J Med Chem*, 2002. **45**(11): p. 2213-21.
83. Cummings, M.D., et al., Comparison of automated docking programs as virtual screening tools. *J Med Chem*, 2005. **48**(4): p. 962-76.
84. Wang, J.L., et al., 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): p. 7124-9.
85. Evers, A. and T. Klabunde, Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor. *J Med Chem*, 2005. **48**(4): p. 1088-97.
86. Stiefl, N. and A. Zaliani, A knowledge-based weighting approach to ligand-based virtual screening. *J Chem Inf Model*, 2006. **46**(2): p. 587-96.
87. Lorber, D.M. and B.K. Shoichet, Hierarchical docking of databases of multiple ligand conformations. *Curr Top Med Chem*, 2005. **5**(8): p. 739-49.
88. Pirard, B., J. Brendel, and S. Peukert, The discovery of Kv1.5 blockers as a case study for the application of virtual screening approaches. *J Chem Inf Model*, 2005. **45**(2): p. 477-85.
89. Rella, M., et al., Structure-based pharmacophore design and virtual screening for novel angiotensin converting enzyme 2 inhibitors. *J Chem Inf Model*, 2006. **46**(2): p. 708-16.
90. Zeng, C. and P.A. Jose, Dopamine receptors: important antihypertensive counterbalance against hypertensive factors. *Hypertension*, 2011. **57**(1): p. 11-7.
91. Cho, D.I., M. Zheng, and K.M. Kim, Current perspectives on the selective regulation of dopamine D2 and D3 receptors. *Arch Pharm Res*, 2010. **33**(10): p. 1521-38.
92. Zhang, A., J.L. Neumeyer, and R.J. Baldessarini, Recent progress in development of dopamine receptor subtype-selective agents: potential therapeutics for neurological and psychiatric disorders. *Chem Rev*, 2007. **107**(1): p. 274-302.
93. Lober, S., et al., Recent advances in the search for D3- and D4-selective drugs: probes, models and candidates. *Trends Pharmacol Sci*, 2011. **32**(3): p. 148-57.
94. Micheli, F., Recent advances in the development of dopamine D3 receptor antagonists: a medicinal chemistry perspective. *ChemMedChem*, 2011. **6**(7): p. 1152-62.
95. Heidbreder, C.A. and A.H. Newman, Current perspectives on selective dopamine D(3) receptor antagonists as pharmacotherapeutics for addictions and related disorders. *Ann N Y Acad Sci*, 2010. **1187**: p. 4-34.
96. Chien, E.Y., et al., Structure of the human dopamine D3 receptor in complex with a D2/D3 selective antagonist. *Science*, 2010. **330**(6007): p. 1091-5.
97. Michielan, L., et al., Comparison of multilabel and single-label classification applied to the prediction of the isoform specificity of cytochrome p450 substrates. *J Chem Inf Model*, 2009. **49**(11): p. 2588-605.
98. Mishra, N.K., S. Agarwal, and G.P. Raghava, Prediction of cytochrome P450 isoform responsible for metabolizing a drug molecule. *BMC Pharmacol*, 2010. **10**: p. 8.
99. Ma, X.H., et al., Virtual Screening of Selective Multitarget Kinase Inhibitors by Combinatorial Support Vector Machines. *Mol Pharm*, 2010. **7**(5): p. 1545-1560.
100. Baek, S., C.A. Tsai, and J.J. Chen, Development of biomarker classifiers from high-dimensional data. *Brief Bioinform*, 2009. **10**(5): p. 537-46.
101. Alizadeh, A.A., et al., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 2000. **403**(6769): p. 503-11.
102. Gordon, G.J., et al., Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res*, 2002. **62**(17): p. 4963-7.
103. Massion, P.P. and D.P. Carbone, The molecular basis of lung cancer: molecular abnormalities and therapeutic implications. *Respir Res*, 2003. **4**: p. 12.
104. Golub, T.R., et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 1999. **286**(5439): p. 531-7.
105. Ramaswamy, S., et al., Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A*, 2001. **98**(26): p. 15149-54.
106. Khan, J., et al., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med*, 2001. **7**(6): p. 673-9.
107. Ross, M.E., et al., Classification of pediatric acute lymphoblastic leukemia by gene

expression profiling. *Blood*, 2003. **102**(8): p. 2951-9.
108. Yeoh, E.J., et al., Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 2002. **1**(2): p. 133-43.
109. Schoch C, D.M., Kern W, Kohlmann A, Schnittger S, Haferlach T, "Deep insight" into microarray technology. *Atlas Genet Cytogenet Oncol Haematol*, 2004.
110. DeRisi, J.L., V.R. Iyer, and P.O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 1997. **278**(5338): p. 680-6.
111. Tavazoie, S., et al., Systematic determination of genetic network architecture. *Nat Genet*, 1999. **22**(3): p. 281-5.
112. Jansen, R., D. Greenbaum, and M. Gerstein, Relating whole-genome expression data with protein-protein interactions. *Genome Res*, 2002. **12**(1): p. 37-46.
113. Ramirez-Benitez Mdel, C., G. Moreno-Hagelsieb, and J.C. Almagro, VIR.II: a new interface with the antibody sequences in the Kabat database. *Biosystems*, 2001. **61**(2-3): p. 125-31.
114. Alon, U., et al., Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A*, 1999. **96**(12): p. 6745-50.
115. Eisen, M.B., et al., Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 1998. **95**(25): p. 14863-8.
116. Sherlock, G., Analysis of large-scale gene expression data. *Curr Opin Immunol*, 2000. **12**(2): p. 201-5.
117. Vapnik, V., Statistical Learning Theory. 1998.
118. Bishop, C., neural networks for pattern recognition. 1995.
119. Inza, I., et al., Filter versus wrapper gene selection approaches in DNA microarray domains. *Artif Intell Med*, 2004. **31**(2): p. 91-103.
120. Model, F., et al., Feature selection for DNA methylation based cancer classification. *Bioinformatics*, 2001. **17 Suppl 1**: p. S157-64.
121. Robnik-Šikonja, M. and I. Kononenko, Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning*, 2003. **53**(1-2): p. 23-69.
122. Ding, C. and H. Peng, Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol*, 2005. **3**(2): p. 185-205.
123. Ben-Bassat, M., Pattern recognition and reduction of dimensionality. *Handbook of statistics II*, 1982: p. p. 773—91.
124. Kohavi, R. and G.H. John, Wrappers for feature subset selection. *Artificial Intelligence*, 97 **Special issue on relevance**(1-2): p. 273 - 324
125. Overington, J., ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. *J Comput Aided Mol Des*, 2009. **23**(4): p. 195-8.
126. Scior, T., et al., How to recognize and workaround pitfalls in QSAR studies: a critical review. *Curr Med Chem*, 2009. **16**(32): p. 4297-313.
127. Susnow, R.G. and S.L. Dixon, Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition. *J Chem Inf Comput Sci*, 2003. **43**(4): p. 1308-15.
128. Perez, J.J., Managing molecular diversity, in *Chemical Society Reviews*2005, *Royal Society of Chemistry*. p. 143-152.
129. Willett, P., J.M. Barnard, and G.M. Downs, Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.*, 1998. **38**(6): p. 983-996.
130. Fang, H., et al., Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens. *Chem Res Toxicol*, 2001. **14**(3): p. 280-94.
131. Tong, W., et al., Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. *Environ Health Perspect*, 2004. **112**(12): p. 1249-54.
132. Hu, J.Y. and T. Aizawa, Quantitative structure-activity relationships for estrogen receptor binding affinity of phenolic chemicals. *Water Res*, 2003. **37**(6): p. 1213-22.
133. Jacobs, M.N., In silico tools to aid risk assessment of endocrine disrupting chemicals. *Toxicology*, 2004. **205**(1-2): p. 43-53.
134. Byvatov, E., et al., Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J Chem Inf Comput Sci*, 2003. **43**(6): p. 1882-9.
135. Doniger, S., T. Hofmann, and J. Yeh, Predicting CNS permeability of drug molecules:

comparison of neural network and support vector machine algorithms. *J Comput Biol*, 2002. **9**(6): p. 849-64.
136. He, L., et al., Predicting the genotoxicity of polycyclic aromatic compounds from molecular structure with different classifiers. *Chem Res Toxicol*, 2003. **16**(12): p. 1567-80.
137. Snyder, R.D., et al., Assessment of the sensitivity of the computational programs DEREK, TOPKAT, and MCASE in the prediction of the genotoxicity of pharmaceutical molecules. *Environ Mol Mutagen*, 2004. **43**(3): p. 143-58.
138. Xue, Y., et al., 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): p. 1630-8.
139. Yap, C.W., et al., Prediction of torsade-causing potential of drugs by support vector machine approach. *Toxicol Sci*, 2004. **79**(1): p. 170-7.
140. Yap, C.W. and Y.Z. Chen, Quantitative Structure-Pharmacokinetic Relationships for drug distribution properties by using general regression neural network. *J Pharm Sci*, 2005. **94**(1): p. 153-68.
141. Yap, C.W., PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem*, 2011. **32**(7): p. 1466-74.
142. Hirsch, F.R., et al., Increased epidermal growth factor receptor gene copy number detected by fluorescence in situ hybridization associates with increased sensitivity to gefitinib in patients with bronchioloalveolar carcinoma subtypes: a Southwest Oncology Group Study. *J Clin Oncol*, 2005. **23**(28): p. 6838-45.
143. Hall LH, K.G., Haney DN, *Molconn-Z*2002: eduSoft LC: Ashland VA.
144. Li, Z.R., et al., MODEL - Molecular descriptor lab: A web-based server for computing structural and physicochemical features of compounds. *Biotechnology and Bioengineering*, 2007. **97**(2): p. 389-396.
145. Steinbeck, C., et al., The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J Chem Inf Comput Sci*, 2003. **43**(2): p. 493-500.
146. Steinbeck, C., et al., Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Curr Pharm Des*, 2006. **12**(17): p. 2111-20.
147. Wegner, J.K., JOELib/JOELib2, 2005: Department of Computer Science,University of Tübingen: Germany.
148. Hemmer, M.C., V. Steinhauer, and J. Gasteiger, Deriving the 3D structure of organic molecules from their infrared spectra. *Vibrational Spectroscopy*, 1999. **19**(1): p. 151-164.
149. Rücker, G. and C. Rücker, Counts of all walks as atomic and molecular descriptors. *Journal of Chemical Information and Computer Sciences*, 1993. **33**(5): p. 683-695.
150. Schuur, J.H., P. Setzer, and J. Gasteiger, The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *Journal of Chemical Information and Computer Sciences*, 1996. **36**(2): p. 334-344.
151. Pearlman, R.S. and K.M. Smith, Metric validation and the receptor-relevant subspace concept. *Journal of Chemical Information and Computer Sciences*, 1999. **39**(1): p. 28-35.
152. Bravi, G., et al., 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): p. 79-92.
153. Galvez, J., et al., Charge indexes. New topological descriptors. *Journal of Chemical Information and Computer Sciences*, 1994. **34**(3): p. 520-525.
154. Consonni, V., R. Todeschini, and M. Pavan, Structure/Response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *Journal of Chemical Information and Computer Sciences*, 2002. **42**(3): p. 682-692.
155. Randic, M., Graph theoretical approach to local and overall aromaticity of benzenoid hydrocarbons. *Tetrahedron*, 1975. **31**(11-12): p. 1477-1481.
156. Randic, M., Molecular profiles. Novel geometry-dependent molecular descriptors. *New Journal of Chemistry*, 1995. **19**: p. 781-791.
157. Kier, L.B. and L.H. Hall, *Molecular structure description: The electrotopological state*1999, San Diego: Academic Press.
158. Platts, J.A., et al., Estimation of molecular free energy relation descriptors using a group contribution approach. *Journal of Chemical Information and Computer Sciences*, 1999. **39**(5): p. 835-845.

159. Liu, T., et al., BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res*, 2007. **35**(Database issue): p. D198-201.
160. Sadowski, J., J. Gasteiger, and G. Klebe, Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.*, 1994. **34**: p. 1000-1008.
161. Dutta, D., et al., Scalable partitioning and exploration of chemical spaces using geometric hashing. *J Chem Inf Model*, 2006. **46**(1): p. 321-33.
162. Parsons, H.M., et al., Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation. *BMC Bioinformatics*, 2007. **8**: p. 234.
163. van den Berg, R.A., et al., Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 2006. **7**: p. 142.
164. Vapnik, V.N., *The nature of statistical learning theory*1995, New York: Springer.
165. Burges, C.J.C., A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 1998. **2**(2): p. 127-167.
166. Pochet, N., et al., Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics*, 2004. **20**: p. 3185-3195.
167. Li, F. and Y. Yang, Analysis of recursive gene selection approaches from microarray data. *Bioinformatics*, 2005. **21**: p. 3741-3747.
168. Jorissen, R.N. and M.K. Gilson, Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model*, 2005. **45**(3): p. 549-61.
169. Glick, M., et al., 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): p. 193-200.
170. Lepp, Z., T. Kinoshita, and H. Chuman, Screening for new antidepressant leads of multiple activities by support vector machines. *J. Chem. Inf. Model*, 2006. **46**(1): p. 158-67.
171. Hert, J., et al., 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): p. 462-70.
172. J. Cui, L.Y.H., 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**: p. 866-877.
173. Yap, C.W. and Y.Z. Chen, Quantitative Structure-Pharmacokinetic Relationships for drug distribution properties by using general regression neural network. *J. Pharm. Sci*, 2005. **94**(1): p. 153-68.
174. Yap, C.W. and Y.Z. Chen, Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J. Chem. Inf. Model*, 2005. **45**(4): p. 982-92.
175. Grover, I.I., I.I. Singh, and I.I. Bakshi, Quantitative structure-property relationships in pharmaceutical research - Part 2. *Pharm. Sci. Technol. Today*, 2000. **3**(2): p. 50-57.
176. Trotter, M.W.B., B.F. Buxton, and S.B. Holden, Support vector machines in combinatorial chemistry. *Meas. Control*, 2001. **34**(8): p. 235-239.
177. Burbidge, R., et al., Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.*, 2001. **26**(1): p. 5-14.
178. Czerminski, R., A. Yasri, and D. Hartsough, Use of support vector machine in pattern classification: Application to QSAR studies. *Quantitative Structure-Activity Relationships*, 2001. **20**(3): p. 227-240.
179. Chang, C.C. and C.J. Lin. *LIBSVM : a library for support vector machines*. 2001; Available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm.
180. Johnson, R.A. and D.W. Wichern, *Applied multivariate statistical analysis*1982, Englewood Cliffs, NJ: Prentice Hall.
181. Fix, E. and J.L. Hodges, *Discriminatory analysis: Non-parametric discrimination: Consistency properties*1951, Texas: USAF School of Aviation Medicine. 261-279.
182. Fujishima, S. and Y. Takahashi, Classification of dopamine antagonists using TFS-based artificial neural network. *J Chem Inf Comput Sci*, 2004. **44**(3): p. 1006-9.
183. Bostrom, J., A. Hogner, and S. Schmitt, Do structurally similar ligands bind in a similar fashion? *J. Med. Chem*, 2006. **49**(23): p. 6716-25.
184. Huang, N., B.K. Shoichet, and J.J. Irwin, Benchmarking sets for molecular docking. *J. Med. Chem*, 2006. **49**(23): p. 6789-801.
185. Matthews, B., Comparison of the predicted and observed secondary structure of T4 phage

lysozyme. *Biochim Biophys Acta*, 1975. **405**(2): p. 442-51.
186. Hawkins, D.M., The problem of overfitting. *J Chem Inf Comput Sci*, 2004. **44**(1): p. 1-12.
187. Furey, T.S., et al., Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 2000. **16**(10): p. 906-14.
188. Leung, Y.F. and D. Cavalieri, Fundamentals of cDNA microarray data analysis. *Trends Genet*, 2003. **19**(11): p. 649-59.
189. Lee, P.D., et al., Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. *Genome Res*, 2002. **12**(2): p. 292-7.
190. Norman Morrison, M.R., Martin Brutsche, Stephen G. Oliver, Andrew Hayes, Nianshu Zhang, Chris Penkett, Jacqui Lockey, Sudha Rao, Ian Hayes, Ray Jupp, Andy Brass, Robust normalization of microarray data over multiple experiments. *Nature Genetics*, 1999. **23**: p. 64.
191. Chu, W., et al., Biomarker discovery in microarray gene expression data with Gaussian processes. *Bioinformatics*, 2005. **21**(16): p. 3385-93.
192. Michael E. Wall, A.R., Luis M. Rocha, *Microarray analysis techniques:Singular value decomposition and principal component analysis*. Understanding and Using Microarray Analysis Techniques: A Practical Guide, ed. W.D. D.P. Berrar, M. Granzow2002: Kluwer Academic Press.
193. Guyon, I., et al., Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 2002. **46**(1-3): p. 389-422.
194. Sima, C., U. Braga-Neto, and E.R. Dougherty, Superior feature-set ranking for small samples using bolstered error estimation. *Bioinformatics*, 2005. **21**(7): p. 1046-54.
195. Fu, W.J., R.J. Carroll, and S. Wang, Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics*, 2005. **21**(9): p. 1979-86.
196. Sibley, D.R. and F.J. Monsma, Jr., Molecular biology of dopamine receptors. *Trends Pharmacol Sci*, 1992. **13**(2): p. 61-9.
197. Simpson, M.M., et al., Dopamine D4/D2 receptor selectivity is determined by A divergent aromatic microdomain contained within the second, third, and seventh membrane-spanning segments. *Mol Pharmacol*, 1999. **56**(6): p. 1116-26.
198. Wang, Q., et al., Subtype selectivity of dopamine receptor ligands: insights from structure and ligand-based methods. *J Chem Inf Model*, 2010. **50**(11): p. 1970-85.
199. Lopez, L., et al., Synthesis, 3D-QSAR, and structural modeling of benzolactam derivatives with binding affinity for the D(2) and D(3) receptors. *ChemMedChem*, 2010. **5**(8): p. 1300-17.
200. Cha, M.Y., et al., QSAR studies on piperazinylalkylisoxazole analogues selectively acting on dopamine D3 receptor by HQSAR and CoMFA. *Bioorg Med Chem*, 2003. **11**(7): p. 1293-8.
201. Audouze, K., E.O. Nielsen, and D. Peters, New series of morpholine and 1,4-oxazepane derivatives as dopamine D4 receptor ligands: synthesis and 3D-QSAR model. *J Med Chem*, 2004. **47**(12): p. 3089-104.
202. Clark, R.D. and E. Abrahamian, Using a staged multi-objective optimization approach to find selective pharmacophore models. *J Comput Aided Mol Des*, 2009. **23**(11): p. 765-71.
203. Salama, I., et al., Structure-selectivity investigations of D2-like receptor ligands by CoMFA and CoMSIA guiding the discovery of D3 selective PET radioligands. *J Med Chem*, 2007. **50**(3): p. 489-500.
204. Carro, L., et al., Synthesis and binding affinity of potential atypical antipsychotics with the tetrahydroquinazolinone motif. *Bioorg Med Chem Lett*, 2009. **19**(21): p. 6059-62.
205. Huber, D., H. Hubner, and P. Gmeiner, 1,1'-Disubstituted ferrocenes as molecular hinges in mono- and bivalent dopamine receptor ligands. *J Med Chem*, 2009. **52**(21): p. 6860-70.
206. Han, L.Y., et al., 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): p. 1276-86.
207. Li, H., et al., Machine learning approaches for predicting compounds that interact with therapeutic and ADMET related proteins. *J Pharm Sci*, 2007. **96**(11): p. 2838-60.
208. Mahe, P. and J.P. Vert, Virtual screening with support vector machines and structure kernels. *Comb Chem High Throughput Screen*, 2009. **12**(4): p. 409-23.
209. Monge, A., et al., Managing, profiling and analyzing a library of 2.6 million compounds gathered from 32 chemical providers. *Molecular diversity*, 2006. **10**(3): p. 389-403.
210. Wang, Y., et al., PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res*, 2009. **37**(Web Server issue): p. W623-33.

211. Irwin, J.J. and B.K. Shoichet, ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model*, 2005. **45**(1): p. 177-82.
212. Bellis, L.J., et al., Collation and data-mining of literature bioactivity data for drug discovery. *Biochemical Society transactions*, 2011. **39**(5): p. 1365-70.
213. Wester, M.J., et al., Scaffold topologies. 2. Analysis of chemical databases. *J Chem Inf Model*, 2008. **48**(7): p. 1311-24.
214. Verheij, H.J., Leadlikeness and structural diversity of synthetic screening libraries. *Molecular diversity*, 2006. **10**(3): p. 377-88.
215. Tsoumakas, G., I. Katakis, and I. Vlahavas, Mining Multi-label Data, in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Editors. 2010, *Springer US*. p. 667-685.
216. Zhang, M.-L. and Z.-H. Zhou, ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 2007. **40**(7): p. 2038-2048.
217. Tsoumakas, G. and I. Katakis, Multi-label classification: An overview. *Int J Data Warehousing and Mining*, 2007. **2007**: p. 1-13.
218. Schietgat, L., et al., Predicting gene function using hierarchical multi-label decision tree ensembles. *BMC Bioinformatics*, 2010. **11**: p. 2.
219. Jenner, P., Dopamine agonists, receptor selectivity and dyskinesia induction in Parkinson's disease. *Curr Opin Neurol*, 2003. **16 Suppl 1**: p. S3-7.
220. McCall, R.B., et al., Sumanirole, a highly dopamine D2-selective receptor agonist: in vitro and in vivo pharmacological characterization and efficacy in animal models of Parkinson's disease. *J Pharmacol Exp Ther*, 2005. **314**(3): p. 1248-56.
221. Singer, C., et al., A comparison of sumanirole versus placebo or ropinirole for the treatment of patients with early Parkinson's disease. *Mov Disord*, 2007. **22**(4): p. 476-82.
222. Pilla, M., et al., Selective inhibition of cocaine-seeking behaviour by a partial dopamine D3 receptor agonist. *Nature*, 1999. **400**(6742): p. 371-5.
223. Boeckler, F. and P. Gmeiner, The structural evolution of dopamine D3 receptor ligands: structure-activity relationships and selected neuropharmacological aspects. *Pharmacol Ther*, 2006. **112**(1): p. 281-333.
224. Albersen, M., et al., The future is today: emerging drugs for the treatment of erectile dysfunction. *Expert Opin Emerg Drugs*, 2010. **15**(3): p. 467-80.
225. Lober, S., et al., The azulene framework as a novel arene bioisostere: design of potent dopamine D4 receptor ligands inducing penile erection. *ChemMedChem*, 2009. **4**(3): p. 325-8.
226. Zhang, J., et al., Dopamine D1 receptor ligands: where are we now and where are we going. *Med Res Rev*, 2009. **29**(2): p. 272-94.
227. Aloisi, G., et al., Differential induction of adenylyl cyclase supersensitivity by antiparkinson drugs acting as agonists at dopamine D1/D2/D3 receptors vs D2/D3 receptors only: parallel observations from co-transfected human and native cerebral receptors. *Neuropharmacology*, 2011. **60**(2-3): p. 439-45.
228. Herm, L., et al., N-Substituted-2-alkyl- and 2-arylnorapomorphines: novel, highly active D2 agonists. *Bioorg Med Chem*, 2009. **17**(13): p. 4756-62.
229. Enguehard-Gueiffier, C. and A. Gueiffier, Recent progress in medicinal chemistry of D4 agonists. *Curr Med Chem*, 2006. **13**(25): p. 2981-93.
230. Ehrlich, K., et al., Dopamine D2, D3, and D4 selective phenylpiperazines as molecular probes to explore the origins of subtype specific receptor binding. *J Med Chem*, 2009. **52**(15): p. 4923-35.
231. Li, Z.R., et al., MODEL-molecular descriptor lab: a web-based server for computing structural and physicochemical features of compounds. *Biotechnol Bioeng*, 2007. **97**(2): p. 389-96.
232. Shi, Z., et al., Combinatorial support vector machines approach for virtual screening of selective multi-target serotonin reuptake inhibitors from large compound libraries. *J Mol Graph Model*, 2012. **32**: p. 49-66.
233. Quinlan, J.R., *C4.5 : programs for machine learning*1993, San Mateo, Calif: Morgan Kaufmann.
234. Matthews, B.W., Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 1975. **405**(2): p. 442-451.
235. Willett, P., Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci*, 1998. **38**: p. 983-996.
236. Kohavi, R. and G.H. John, Wrappers for feature subset selection. *Artificial Intelligence*, 1997. **97**(1–2): p. 273-324.

237. Durrant, J.D. and J.A. McCammon, Computer-aided drug-discovery techniques that account for receptor flexibility. *Curr Opin Pharmacol*, 2010. **10**(6): p. 770-4.
238. Sprous, D.G., et al., QSAR in the pharmaceutical research setting: QSAR models for broad, large problems. *Curr Top Med Chem*, 2010. **10**(6): p. 619-37.
239. Willett, P., Similarity searching using 2D structural fingerprints. *Methods Mol Biol*, 2011. **672**: p. 133-58.
240. Ma, X.H., et al., Comparative analysis of machine learning methods in ligand-based virtual screening of large compound libraries. *Comb Chem High Throughput Screen*, 2009. **12**(4): p. 344-57.
241. Sato, T., T. Honma, and S. Yokoyama, Combining machine learning and pharmacophore-based interaction fingerprint for in silico screening. *J Chem Inf Model*, 2010. **50**(1): p. 170-85.
242. Talevi, A., L. Gavernet, and L.E. Bruno-Blanch, Combined Virtual Screening Strategies. *Current Computer - Aided Drug Design*, 2009. **5**(1): p. 23-37.
243. Bender, A., et al., "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): p. 2445-56.
244. Givehchi, A., A. Bender, and R.C. Glen, Analysis of activity space by fragment fingerprints, 2D descriptors, and multitarget dependent transformation of 2D descriptors. *J Chem Inf Model*, 2006. **46**(3): p. 1078-83.
245. Renner, S., et al., Maximum common binding modes (MCBM): consensus docking scoring using multiple ligand information and interaction fingerprints. *J Chem Inf Model*, 2008. **48**(2): p. 319-32.
246. Erhan, D., et al., Collaborative filtering on a family of biological targets. *J Chem Inf Model*, 2006. **46**(2): p. 626-35.
247. Dragos, H., M. Gilles, and V. Alexandre, Predicting the predictability: a unified approach to the applicability domain problem of QSAR models. *J Chem Inf Model*, 2009. **49**(7): p. 1762-76.
248. Liu, X.H., et al., Virtual screening of Abl inhibitors from large compound libraries by support vector machines. *J Chem Inf Model*, 2009. **49**(9): p. 2101-10.
249. Greten, F.R., et al., IKKbeta links inflammation and tumorigenesis in a mouse model of colitis-associated cancer. *Cell*, 2004. **118**(3): p. 285-96.
250. Ghosh, S. and M.S. Hayden, New regulators of NF-kappaB in inflammation. *Nat Rev Immunol*, 2008. **8**(11): p. 837-48.
251. Strnad, J. and J.R. Burke, IkappaB kinase inhibitors for treating autoimmune and inflammatory disorders: potential and challenges. *Trends Pharmacol Sci*, 2007. **28**(3): p. 142-8.
252. Zhang, Y., et al., IkappaBalpha kinase inhibitor IKI-1 conferred tumor necrosis factor alpha sensitivity to pancreatic cancer cells and a xenograft tumor model. *Cancer Res*, 2008. **68**(22): p. 9519-24.
253. Schon, M., et al., KINK-1, a novel small-molecule inhibitor of IKKbeta, and the susceptibility of melanoma cells to antitumoral treatment. *J Natl Cancer Inst*, 2008. **100**(12): p. 862-75.
254. Ma, X.H., et al., Evaluation of virtual screening performance of support vector machines trained by sparsely distributed active compounds. *J Chem Inf Model*, 2008. **48**(6): p. 1227-37.
255. Bocker, A., G. Schneider, and A. Teckentrup, NIPALSTREE: a new hierarchical clustering approach for large compound libraries and its application to virtual screening. *J. Chem. Inf. Model*, 2006. **46**(6): p. 2220-9.
256. Oprea, T.I. and J. Gottfries, Chemography: the art of navigating in chemical space. *J. Comb. Chem*, 2001. **3**(2): p. 157-66.
257. Fang, H., et al., Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens. *Chem. Res. Toxicol.*, 2001. **14**: p. 280-294.
258. Tong, W., et al., Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. *Environ. Health Perspect.*, 2004. **112**(12): p. 1249-1254.
259. Hu, J.Y. and T. Aizawa, Quantitative structure-activity relationships for estrogen receptor binding affinity of phenolic chemicals. *Water Res.*, 2003. **37**(6): p. 1213-1222.
260. Jacobs, M.N., In silico tools to aid risk assessment of endocrine disrupting chemicals. *Toxicology*, 2004. **205**(1-2): p. 43-53.

261. Byvatov, E., et al., Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci*, 2003. **43**(6): p. 1882-1889.
262. Doniger, S., T. Hofman, and J. Yeh, Predicting CNS Permeability of Drug Molecules:Comparison of Neural Network and Support Vector Machine Algorithms. *J. Comput. Biol.*, 2002. **9**(6): p. 849-864.
263. He, L., et al., Predicting the Genotoxicity of Polycyclic Aromatic Compounds from Molecular Structure with Different Classifiers. *Chem. Res. Toxicol.*, 2003. **16**(12): p. 1567-1580.
264. Snyder, R.D., et al., Assessment of the sensitivity of the computational programs DEREK, TOPKAT, and MCASE in the prediction of the genotoxicity of pharmaceutical molecules. *Environ. Mol. Mutagen.*, 2004. **43**(3): p. 143-158.
265. Xue, Y., et al., 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): p. 1630-1638.
266. Xue, Y., et al., Prediction of P-glycoprotein substrates by a support vector machine approach. *J. Chem. Inf. Comput. Sci*, 2004. **44**(4): p. 1497-505.
267. Mayer, D., F. Leisch, and K. Hornik, The support vector machine under test. *Neurocomputing*, 2003. **55**(1-2): p. 169-186.
268. Ho, C. and J. Laskin, EGFR-directed therapies to treat non-small-cell lung cancer. *Expert Opin Investig Drugs*, 2009. **18**(8): p. 1133-45.
269. Linardou, H., et al., Somatic EGFR mutations and efficacy of tyrosine kinase inhibitors in NSCLC. *Nat Rev Clin Oncol*, 2009. **6**(6): p. 352-66.
270. Paez, J.G., et al., EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science*, 2004. **304**(5676): p. 1497-500.
271. Sharma, S.V., et al., Epidermal growth factor receptor mutations in lung cancer. *Nat Rev Cancer*, 2007. **7**(3): p. 169-81.
272. Stemke-Hale, K., et al., An integrative genomic and proteomic analysis of PIK3CA, PTEN, and AKT mutations in breast cancer. *Cancer Res*, 2008. **68**(15): p. 6084-91.
273. Vivanco, I. and C.L. Sawyers, The phosphatidylinositol 3-Kinase AKT pathway in human cancer. *Nat Rev Cancer*, 2002. **2**(7): p. 489-501.
274. Laurent-Puig, P., A. Lievre, and H. Blons, Mutations and response to epidermal growth factor receptor inhibitors. *Clin Cancer Res*, 2009. **15**(4): p. 1133-9.
275. Sartore-Bianchi, A., et al., PIK3CA mutations in colorectal cancer are associated with clinical resistance to EGFR-targeted monoclonal antibodies. *Cancer Res*, 2009. **69**(5): p. 1851-7.
276. Teng, D.H., et al., MMAC1/PTEN mutations in primary tumor specimens and tumor cell lines. *Cancer Res*, 1997. **57**(23): p. 5221-5.
277. Bianco, R., et al., Loss of PTEN/MMAC1/TEP in EGF receptor-expressing tumor cells counteracts the antitumor action of EGFR tyrosine kinase inhibitors. *Oncogene*, 2003. **22**(18): p. 2812-22.
278. Mellinghoff, I.K., T.F. Cloughesy, and P.S. Mischel, PTEN-mediated resistance to epidermal growth factor receptor kinase inhibitors. *Clin Cancer Res*, 2007. **13**(2 Pt 1): p. 378-81.
279. Sequist, L.V., et al., Molecular predictors of response to epidermal growth factor receptor antagonists in non-small-cell lung cancer. *J Clin Oncol*, 2007. **25**(5): p. 587-95.
280. Sos, M.L., et al., Predicting drug susceptibility of non-small cell lung cancers based on genetic lesions. *J Clin Invest*, 2009. **119**(6): p. 1727-40.
281. Emery, I.F., et al., Response to gefitinib and erlotinib in Non-small cell lung cancer: a restrospective study. *BMC Cancer*, 2009. **9**: p. 333.
282. Petak, I., et al., Integrating molecular diagnostics into anticancer drug discovery. *Nat Rev Drug Discov*. **9**(7): p. 523-35.
283. Barton, S., N. Starling, and C. Swanton, Predictive Molecular Markers of Response to Epidermal Growth Factor Receptor(EGFR) Family-Targeted Therapies. *Curr Cancer Drug Targets*, 2010. **10**(8): p. 799-812.
284. Roberts, P.J., et al., Personalized medicine in non-small-cell lung cancer: is KRAS a useful marker in selecting patients for epidermal growth factor receptor-targeted therapy? *J Clin Oncol*. **28**(31): p. 4769-77.
285. Sergina, N.V., et al., Escape from HER-family tyrosine kinase inhibitor therapy by the kinase-inactive HER3. *Nature*, 2007. **445**(7126): p. 437-41.

286. Sawyers, C.L., Cancer: mixing cocktails. *Nature*, 2007. **449**(7165): p. 993-6.
287. Jia, J., et al., Mechanisms of drug combinations: interaction and network perspectives. *Nat Rev Drug Discov*, 2009. **8**(2): p. 111-28.
288. Knight, Z.A., H. Lin, and K.M. Shokat, Targeting the cancer kinome through polypharmacology. *Nat Rev Cancer*. **10**(2): p. 130-7.
289. Gossage, L. and T. Eisen, Targeting multiple kinase pathways: a change in paradigm. *Clin Cancer Res*. **16**(7): p. 1973-8.
290. Wheeler, D.L., E.F. Dunn, and P.M. Harari, Understanding resistance to EGFR inhibitors-impact on future treatment strategies. *Nat Rev Clin Oncol*. **7**(9): p. 493-507.
291. Erjala, K., et al., Signaling via ErbB2 and ErbB3 associates with resistance and epidermal growth factor receptor (EGFR) amplification with sensitivity to EGFR inhibitor gefitinib in head and neck squamous cell carcinoma cells. *Clin Cancer Res*, 2006. **12**(13): p. 4103-11.
292. Oshita, F., et al., Genomic-wide cDNA microarray screening to correlate gene expression profile with chemoresistance in patients with advanced lung cancer. *J Exp Ther Oncol*, 2004. **4**(2): p. 155-60.
293. Coldren, C.D., et al., Baseline gene expression predicts sensitivity to gefitinib in non-small cell lung cancer cell lines. *Mol Cancer Res*, 2006. **4**(8): p. 521-8.
294. Vegran, F., et al., Gene expression profile and response to trastuzumab-docetaxel-based treatment in breast carcinoma. *Br J Cancer*, 2009. **101**(8): p. 1357-64.
295. Okano, T., et al., Proteomic signature corresponding to the response to gefitinib (Iressa, ZD1839), an epidermal growth factor receptor tyrosine kinase inhibitor in lung adenocarcinoma. *Clin Cancer Res*, 2007. **13**(3): p. 799-805.
296. Winegarden, N., Microarrays in cancer: moving from hype to clinical reality. *Lancet*, 2003. **362**(9394): p. 1428.
297. Debouck, C. and P.N. Goodfellow, DNA microarrays in drug discovery and development. *Nat Genet*, 1999. **21**(1 Suppl): p. 48-50.
298. Michiels, S., S. Koscielny, and C. Hill, Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, 2005. **365**(9458): p. 488-92.
299. Bullinger, L. and P.J. Valk, Gene expression profiling in acute myeloid leukemia. *J Clin Oncol*, 2005. **23**(26): p. 6296-305.
300. Allison, D.B., et al., Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet*, 2006. **7**(1): p. 55-65.
301. Tang, Z.Q., et al., Derivation of stable microarray cancer-differentiating signatures using consensus scoring of multiple random sampling and gene-ranking consistency evaluation. *Cancer Res*, 2007. **67**(20): p. 9996-10003.
302. Fan, X., et al., DNA microarrays are predictive of cancer prognosis: a re-evaluation. *Clin Cancer Res*. **16**(2): p. 629-36.
303. Sharma, S.V., D.A. Haber, and J. Settleman, Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nat Rev Cancer*, 2010. **10**(4): p. 241-53.
304. Burchert, A., et al., Compensatory PI3-kinase/Akt/mTor activation regulates imatinib resistance development. *Leukemia*, 2005. **19**(10): p. 1774-82.
305. Engelman, J.A., et al., MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. *Science*, 2007. **316**(5827): p. 1039-43.
306. Montagut, C., et al., Elevated CRAF as a potential mechanism of acquired resistance to BRAF inhibition in melanoma. *Cancer Res*, 2008. **68**(12): p. 4853-61.
307. Cools, J., et al., Prediction of resistance to small molecule FLT3 inhibitors: implications for molecularly targeted therapy of acute leukemia. *Cancer Res*, 2004. **64**(18): p. 6385-9.
308. Papadopoulos, N., K.W. Kinzler, and B. Vogelstein, The role of companion diagnostics in the development and use of mutation-targeted cancer therapies. *Nat Biotechnol*, 2006. **24**(8): p. 985-95.
309. Rizvi, N.A., et al., Molecular characteristics predict clinical outcomes: prospective trial correlating response to the EGFR tyrosine kinase inhibitor gefitinib with the presence of sensitizing mutations in the tyrosine binding domain of the EGFR gene. *Clin Cancer Res*, 2011. **17**(10): p. 3500-6.
310. Molinari, F., et al., Increased detection sensitivity for KRAS mutations enhances the prediction of anti-EGFR monoclonal antibody resistance in metastatic colorectal cancer. *Clin Cancer Res*, 2011. **17**(14): p. 4901-14.
311. Kudoh, K., et al., Gains of 1q21-q22 and 13q12-q14 are potential indicators for resistance to cisplatin-based chemotherapy in ovarian cancer patients. *Clin Cancer Res*, 1999. **5**(9): p. 2526-31.

312. Hilsenbeck, S.G., et al., Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *J Natl Cancer Inst*, 1999. **91**(5): p. 453-9.
313. Rosenwald, A., et al., The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med*, 2002. **346**(25): p. 1937-47.
314. Staunton, J.E., et al., Chemosensitivity prediction by transcriptional profiling. *Proc Natl Acad Sci U S A*, 2001. **98**(19): p. 10787-92.
315. Takata, R., et al., Predicting response to methotrexate, vinblastine, doxorubicin, and cisplatin neoadjuvant chemotherapy for bladder cancers through genome-wide gene expression profiling. *Clin Cancer Res*, 2005. **11**(7): p. 2625-36.
316. Ma, Y., et al., Predicting cancer drug response by proteomic profiling. *Clin Cancer Res*, 2006. **12**(15): p. 4583-9.
317. Tabchy, A., et al., Evaluation of a 30-gene paclitaxel, fluorouracil, doxorubicin, and cyclophosphamide chemotherapy response predictor in a multicenter randomized trial in breast cancer. *Clin Cancer Res*, 2010. **16**(21): p. 5351-61.
318. Hatzis, C., et al., A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA*, 2011. **305**(18): p. 1873-81.
319. Cepero, V., et al., MET and KRAS gene amplification mediates acquired resistance to MET tyrosine kinase inhibitors. *Cancer Res*, 2010. **70**(19): p. 7580-90.
320. Di Leo, A., et al., HER2 and TOP2A as predictive markers for anthracycline-containing chemotherapy regimens as adjuvant treatment of breast cancer: a meta-analysis of individual patient data. *Lancet Oncol*, 2011. **12**(12): p. 1134-42.
321. Coon, J.S., et al., Amplification and overexpression of topoisomerase IIalpha predict response to anthracycline-based therapy in locally advanced breast cancer. *Clin Cancer Res*, 2002. **8**(4): p. 1061-7.
322. Broet, P., et al., Prediction of clinical outcome in multiple lung cancer cohorts by integrative genomics: implications for chemotherapy selection. *Cancer Res*, 2009. **69**(3): p. 1055-62.
323. John, T., G. Liu, and M.S. Tsao, Overview of molecular testing in non-small-cell lung cancer: mutational analysis, gene copy number, protein expression and other biomarkers of EGFR for the prediction of response to tyrosine kinase inhibitors. *Oncogene*, 2009. **28 Suppl 1**: p. S14-23.
324. Tiseo, M., et al., Predictors of gefitinib outcomes in advanced non-small cell lung cancer (NSCLC): study of a comprehensive panel of molecular markers. *Lung Cancer*, 2010. **67**(3): p. 355-60.
325. Wang, W., et al., Mechanistic and predictive profiling of 5-Fluorouracil resistance in human cancer cells. *Cancer Res*, 2004. **64**(22): p. 8167-76.
326. Dai, Z., et al., Prediction of anticancer drug potency from expression of genes involved in growth factor signaling. *Pharm Res*, 2006. **23**(2): p. 336-49.
327. Brase, J.C., et al., ERBB2 and TOP2A in breast cancer: a comprehensive analysis of gene amplification, RNA levels, and protein expression and their influence on prognosis and prediction. *Clin Cancer Res*, 2010. **16**(8): p. 2391-401.
328. Balko, J.M., et al., Gene expression patterns that predict sensitivity to epidermal growth factor receptor tyrosine kinase inhibitors in lung cancer cell lines and human lung tumors. *BMC Genomics*, 2006. **7**: p. 289.
329. Romano, P., et al., Cell Line Data Base: structure and recent improvements towards molecular authentication of human cell lines. *Nucleic Acids Res*, 2009. **37**(Database issue): p. D925-32.
330. Wang, H., et al., Chemical data mining of the NCI human tumor cell line database. *J Chem Inf Model*, 2007. **47**(6): p. 2063-76.
331. Richard, S.M., et al., Nuclear and mitochondrial genome instability in human breast cancer. *Cancer Res*, 2000. **60**(15): p. 4231-7.
332. Calado, R.T., et al., Constitutional hypomorphic telomerase mutations in patients with acute myeloid leukemia. *Proc Natl Acad Sci U S A*, 2009. **106**(4): p. 1187-92.
333. Ross-Innes, C.S., et al., Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, 2012. **481**(7381): p. 389-93.
334. Gillet, J.P., et al., Redefining the relevance of established cancer cell lines to the study of mechanisms of clinical anti-cancer drug resistance. *Proc Natl Acad Sci U S A*, 2011. **108**(46): p. 18708-13.
335. Leung, C.T. and J.S. Brugge, Outgrowth of single oncogene-expressing cells from

suppressive epithelial environments. *Nature*, 2012. **482**(7385): p. 410-3.
336. Pickl, M. and C.H. Ries, Comparison of 3D and 2D tumor models reveals enhanced HER2 activation in 3D associated with an increased response to trastuzumab. *Oncogene*, 2009. **28**(3): p. 461-8.
337. Weigelt, B., et al., HER2 signaling pathway activation and response of breast cancer cells to HER2-targeting agents is dependent strongly on the 3D microenvironment. *Breast Cancer Res Treat*, 2010. **122**(1): p. 35-43.
338. Sierra, J.R., V. Cepero, and S. Giordano, Molecular mechanisms of acquired resistance to tyrosine kinase targeted therapy. *Mol Cancer*. **9**: p. 75.
339. Forgacs, E., et al., Mutation analysis of the PTEN/MMAC1 gene in lung cancer. *Oncogene*, 1998. **17**(12): p. 1557-65.
340. Nevo, J., et al., Mammary-derived growth inhibitor alters traffic of EGFR and induces a novel form of cetuximab resistance. *Clin Cancer Res*, 2009. **15**(21): p. 6570-81.
341. Citri, A. and Y. Yarden, EGF-ERBB signalling: towards the systems level. *Nat Rev Mol Cell Biol*, 2006. **7**(7): p. 505-16.
342. Kong, A., et al., HER2 oncogenic function escapes EGFR tyrosine kinase inhibitors via activation of alternative HER receptors in breast cancer cells. *PLoS One*, 2008. **3**(8): p. e2881.
343. Engelman, J.A. and L.C. Cantley, The role of the ErbB family members in non-small cell lung cancers sensitive to epidermal growth factor receptor kinase inhibitors. *Clin Cancer Res*, 2006. **12**(14 Pt 2): p. 4372s-4376s.
344. Shi, F., et al., ErbB3/HER3 intracellular domain is competent to bind ATP and catalyze autophosphorylation. *Proc Natl Acad Sci U S A*, 2010. **107**(17): p. 7692-7.
345. Morgillo, F., et al., Implication of the insulin-like growth factor-IR pathway in the resistance of non-small cell lung cancer cells to treatment with gefitinib. *Clin Cancer Res*, 2007. **13**(9): p. 2795-803.
346. Guix, M., et al., Acquired resistance to EGFR tyrosine kinase inhibitors in cancer cells is mediated by loss of IGF-binding proteins. *J Clin Invest*, 2008. **118**(7): p. 2609-19.
347. Agarwal, S., et al., Association of constitutively activated hepatocyte growth factor receptor (Met) with resistance to a dual EGFR/Her2 inhibitor in non-small-cell lung cancer cells. *Br J Cancer*, 2009. **100**(6): p. 941-9.
348. Benedettini, E., et al., Met activation in non-small cell lung cancer is associated with de novo resistance to EGFR inhibitors and the development of brain metastasis. *Am J Pathol*, 2010. **177**(1): p. 415-23.
349. Liska, D., et al., HGF rescues colorectal cancer cells from EGFR inhibition via MET activation. *Clin Cancer Res*, 2011. **17**(3): p. 472-82.
350. Thomson, S., et al., Kinase switching in mesenchymal-like non-small cell lung cancer lines contributes to EGFR inhibitor resistance through pathway redundancy. *Clin Exp Metastasis*, 2008. **25**(8): p. 843-54.
351. Kono, S.A., et al., The fibroblast growth factor receptor signaling pathway as a mediator of intrinsic resistance to EGFR-specific tyrosine kinase inhibitors in non-small cell lung cancer. *Drug Resist Updat*, 2009. **12**(4-5): p. 95-102.
352. Naumov, G.N., et al., Combined vascular endothelial growth factor receptor and epidermal growth factor receptor (EGFR) blockade inhibits tumor growth in xenograft models of EGFR inhibitor resistance. *Clin Cancer Res*, 2009. **15**(10): p. 3484-94.
353. Cabodi, S., et al., Convergence of integrins and EGF receptor signaling via PI3K/Akt/FoxO pathway in early gene Egr-1 expression. *J Cell Physiol*, 2009. **218**(2): p. 294-303.
354. Zeller, K.S., et al., PI3-kinase p110alpha mediates beta1 integrin-induced Akt activation and membrane protrusion during cell attachment and initial spreading. *Cell Signal*. **22**(12): p. 1838-48.
355. Ju, L., et al., Integrin beta1 over-expression associates with resistance to tyrosine kinase inhibitor gefitinib in non-small cell lung cancer. *J Cell Biochem*. **111**(6): p. 1565-74.
356. Velling, T., A. Stefansson, and S. Johansson, EGFR and beta1 integrins utilize different signaling pathways to activate Akt. *Exp Cell Res*, 2008. **314**(2): p. 309-16.
357. Yao, Z., et al., TGF-beta IL-6 axis mediates selective and adaptive mechanisms of resistance to molecular targeted therapy in lung cancer. *Proc Natl Acad Sci U S A*, 2010. **107**(35): p. 15535-40.
358. Sriuranpong, V., et al., Epidermal growth factor receptor-independent constitutive activation of STAT3 in head and neck squamous cell carcinoma is mediated by the autocrine/paracrine stimulation of the interleukin 6/gp130 cytokine system. *Cancer Res*,

2003. **63**(11): p. 2948-56.
359. Kim, Y.M., S.Y. Park, and H. Pyo, Cyclooxygenase-2 (COX-2) negatively regulates expression of epidermal growth factor receptor and causes resistance to gefitinib in COX-2-overexpressing cancer cells. *Mol Cancer Res*, 2009. **7**(8): p. 1367-77.
360. Wu, W.K., et al., Cyclooxygenase-2 in tumorigenesis of gastrointestinal cancers: an update on the molecular mechanisms. *Cancer Lett*, 2010. **295**(1): p. 7-16.
361. Raponi, M., H. Winkler, and N.C. Dracopoli, KRAS mutations predict response to EGFR inhibitors. *Curr Opin Pharmacol*, 2008. **8**(4): p. 413-8.
362. Linardou, H., et al., Assessment of somatic k-RAS mutations as a mechanism associated with resistance to EGFR-targeted agents: a systematic review and meta-analysis of studies in advanced non-small-cell lung cancer and metastatic colorectal cancer. *Lancet Oncol*, 2008. **9**(10): p. 962-72.
363. Vivanco, I., et al., The phosphatase and tensin homolog regulates epidermal growth factor receptor (EGFR) inhibitor response by targeting EGFR for degradation. *Proc Natl Acad Sci U S A*, 2010. **107**(14): p. 6459-64.
364. Weinstein, I.B. and A.K. Joe, Mechanisms of disease: Oncogene addiction--a rationale for molecular targeting in cancer therapy. *Nat Clin Pract Oncol*, 2006. **3**(8): p. 448-57.
365. Lynch, T.J., et al., Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med*, 2004. **350**(21): p. 2129-39.
366. Mok, T.S., et al., Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N Engl J Med*, 2009. **361**(10): p. 947-57.
367. Shepherd, F.A., et al., Erlotinib in previously treated non-small-cell lung cancer. *N Engl J Med*, 2005. **353**(2): p. 123-32.
368. Bekaii-Saab, T., et al., A multi-institutional phase II study of the efficacy and tolerability of lapatinib in patients with advanced hepatocellular carcinomas. *Clin Cancer Res*, 2009. **15**(18): p. 5895-901.
369. Ross, H.J., et al., Randomized phase II multicenter trial of two schedules of lapatinib as first- or second-line monotherapy in patients with advanced or metastatic non-small cell lung cancer. *Clin Cancer Res*, 2010. **16**(6): p. 1938-49.
370. Sridhar, S.S., et al., A multicenter phase II clinical trial of lapatinib (GW572016) in hormonally untreated advanced prostate cancer. *Am J Clin Oncol*. **33**(6): p. 609-13.
371. Kakiuchi, S., et al., Prediction of sensitivity of advanced non-small cell lung cancers to gefitinib (Iressa, ZD1839). *Hum Mol Genet*, 2004. **13**(24): p. 3029-43.
372. Yarden, Y. and M.X. Sliwkowski, Untangling the ErbB signalling network. *Nat Rev Mol Cell Biol*, 2001. **2**(2): p. 127-37.
373. Hynes, N.E. and H.A. Lane, ERBB receptors and cancer: the complexity of targeted inhibitors. *Nat Rev Cancer*, 2005. **5**(5): p. 341-54.
374. Normanno, N., et al., Epidermal growth factor receptor (EGFR) signaling in cancer. *Gene*, 2006. **366**(1): p. 2-16.
375. Wheeler, D.L., et al., Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res*, 2004. **32**(Database issue): p. D35-40.
376. McDermott, U., et al., Identification of genotype-correlated sensitivity to selective kinase inhibitors by using high-throughput tumor cell line profiling. *Proc Natl Acad Sci U S A*, 2007. **104**(50): p. 19936-41.
377. Yun, C.H., et al., The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proc Natl Acad Sci U S A*, 2008. **105**(6): p. 2070-5.
378. Bos, J.L., The ras gene family and human carcinogenesis. *Mutat Res*, 1988. **195**(3): p. 255-71.
379. Bader, A.G., S. Kang, and P.K. Vogt, Cancer-specific mutations in PIK3CA are oncogenic in vivo. *Proc Natl Acad Sci U S A*, 2006. **103**(5): p. 1475-9.
380. Gymnopoulos, M., M.A. Elsliger, and P.K. Vogt, Rare cancer-specific mutations in PIK3CA show gain of function. *Proc Natl Acad Sci U S A*, 2007. **104**(13): p. 5569-74.
381. UniProt_Consortium, The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res*, 2009. **37**(Database issue): p. D169-74.
382. Ercan, D., et al., Amplification of EGFR T790M causes resistance to an irreversible EGFR inhibitor. *Oncogene*, 2010. **29**(16): p. 2346-56.
383. Kitazaki, T., et al., Gefitinib, an EGFR tyrosine kinase inhibitor, directly inhibits the function of P-glycoprotein in multidrug resistant cancer cells. *Lung Cancer*, 2005. **49**(3): p. 337-43.

384. Oprea, T.I., et al., A crowdsourcing evaluation of the NIH chemical probes. *Nat Chem Biol*, 2009. **5**(7): p. 441-7.
385. Forbes, S.A., et al., COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res*, 2011. **39**(Database issue): p. D945-50.
386. Barrett, T., et al., NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic Acids Res*, 2011. **39**(Database issue): p. D1005-10.
387. Irizarry, R.A., et al., Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 2003. **4**(2): p. 249-64.
388. Zhou, B.B., et al., Targeting ADAM-mediated ligand cleavage to inhibit HER3 and EGFR pathways in non-small cell lung cancer. *Cancer Cell*, 2006. **10**(1): p. 39-50.
389. Forbes, S.A., et al., COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res*. **38**(Database issue): p. D652-7.
390. Thomas, R.K., et al., High-throughput oncogene mutation profiling in human cancer. *Nat Genet*, 2007. **39**(3): p. 347-51.
391. Amann, J., et al., Aberrant epidermal growth factor receptor signaling and enhanced sensitivity to EGFR inhibitors in lung cancer. *Cancer Res*, 2005. **65**(1): p. 226-35.
392. Gandhi, J., et al., Alterations in genes of the EGFR signaling pathway and their relationship to EGFR tyrosine kinase inhibitor sensitivity in lung cancer cell lines. *PLoS One*, 2009. **4**(2): p. e4576.
393. Kadara, H., et al., Identification of gene signatures and molecular markers for human lung cancer prognosis using an in vitro lung carcinogenesis system. *Cancer Prev Res (Phila Pa)*, 2009. **2**(8): p. 702-11.
394. Eckford, P.D. and F.J. Sharom, ABC efflux pump-based resistance to chemotherapy drugs. *Chem Rev*, 2009. **109**(7): p. 2989-3011.
395. Noguchi, K., et al., Substrate-dependent bidirectional modulation of P-glycoprotein-mediated drug resistance by erlotinib. *Cancer Sci*, 2009. **100**(9): p. 1701-7.
396. Polli, J.W., et al., The role of efflux and uptake transporters in [N-{3-chloro-4-[(3-fluorobenzyl)oxy]phenyl}-6-[5-({[2-(methylsulfonyl)ethy l]amino}methyl)-2-furyl]-4-quinazolinamine (GW572016, lapatinib) disposition and drug interactions. *Drug Metab Dispos*, 2008. **36**(4): p. 695-701.
397. Pollack, J.R., et al., Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A*, 2002. **99**(20): p. 12963-8.
398. Perren, A., et al., Immunohistochemical evidence of loss of PTEN expression in primary ductal adenocarcinomas of the breast. *Am J Pathol*, 1999. **155**(4): p. 1253-60.
399. Li, J., et al., PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science*, 1997. **275**(5308): p. 1943-7.
400. Hynes, N.E. and T. Schlange, Targeting ADAMS and ERBBs in lung cancer. *Cancer Cell*, 2006. **10**(1): p. 7-11.
401. Jonsson, G., et al., High-resolution genomic profiles of breast cancer cell lines assessed by tiling BAC array comparative genomic hybridization. *Genes Chromosomes Cancer*, 2007. **46**(6): p. 543-58.
402. Cappuzzo, F., et al., Epidermal growth factor receptor gene and protein and gefitinib sensitivity in non-small-cell lung cancer. *J Natl Cancer Inst*, 2005. **97**(9): p. 643-55.
403. Redon, R., et al., Global variation in copy number in the human genome. *Nature*, 2006. **444**(7118): p. 444-54.
404. Draghici, S., Statistical intelligence: effective analysis of high-density microarray data. *Drug Discov Today*, 2002. **7**(11 Suppl): p. S55-63.
405. Sequist, L.V., et al., Genotypic and histological evolution of lung cancers acquiring resistance to EGFR inhibitors. *Sci Transl Med*, 2011. **3**(75): p. 75ra26.
406. Gazdar, A.F., Activating and resistance mutations of EGFR in non-small-cell lung cancer: role in clinical response to EGFR tyrosine kinase inhibitors. *Oncogene*, 2009. **28 Suppl 1**: p. S24-31.
407. Sanchez-Palencia, A., et al., Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *Int J Cancer*, 2011. **129**(2): p. 355-64.
408. Isabelle Guyon, J.W., Stephen Barnhill, Vladimir Vapnik, Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 2002. **46**(1-3): p. 389-422.
409. Pochet, N., et al., Systematic benchmarking of microarray data classification: assessing the

role of non-linearity and dimensionality reduction. *Bioinformatics*, 2004. **20**(17): p. 3185-95.
410. Qiu, P., Z.J. Wang, and K.J. Liu, Ensemble dependence model for classification and prediction of cancer and normal gene expression data. *Bioinformatics*, 2005. **21**(14): p. 3114-21.
411. Li, F. and Y. Yang, Analysis of recursive gene selection approaches from microarray data. *Bioinformatics*, 2005. **21**(19): p. 3741-7.
412. Zhou, W., et al., Novel mutant-selective EGFR kinase inhibitors against EGFR T790M. *Nature*, 2009. **462**(7276): p. 1070-4.
413. Olaussen, K.A., et al., Loss of PTEN expression is not uncommon, but lacks prognostic value in stage I NSCLC. *Anticancer Res*, 2003. **23**(6C): p. 4885-90.
414. Olayioye, M.A., et al., The ErbB signaling network: receptor heterodimerization in development and cancer. *EMBO J*, 2000. **19**(13): p. 3159-67.
415. Gschwind, A., et al., Cell communication networks: epidermal growth factor receptor transactivation as the paradigm for interreceptor signal transmission. *Oncogene*, 2001. **20**(13): p. 1594-600.
416. Uramoto, H. and T. Mitsudomi, Which biomarker predicts benefit from EGFR-TKI treatment for patients with lung cancer? *Br J Cancer*, 2007. **96**(6): p. 857-63.
417. Laurent-Puig, P., et al., Analysis of PTEN, BRAF, and EGFR status in determining benefit from cetuximab therapy in wild-type KRAS metastatic colon cancer. *J Clin Oncol*, 2009. **27**(35): p. 5924-30.
418. Eichmann, A., et al., Ligand-dependent development of the endothelial and hemopoietic lineages from embryonic mesodermal cells expressing vascular endothelial growth factor receptor 2. *Proc Natl Acad Sci U S A*, 1997. **94**(10): p. 5141-6.
419. Ohri, C.M., et al., The tissue microlocalisation and cellular expression of CD163, VEGF, HLA-DR, iNOS, and MRP 8/14 is correlated to clinical outcome in NSCLC. *PLoS One*, 2011. **6**(7): p. e21874.
420. Carrillo de Santa Pau, E., et al., Prognostic significance of the expression of vascular endothelial growth factors A, B, C, and D and their receptors R1, R2, and R3 in patients with nonsmall cell lung cancer. *Cancer*, 2009. **115**(8): p. 1701-12.
421. Bonnesen, B., et al., Vascular endothelial growth factor A and vascular endothelial growth factor receptor 2 expression in non-small cell lung cancer patients: relation to prognosis. *Lung Cancer*, 2009. **66**(3): p. 314-8.
422. Jantus-Lewintre, E., et al., Combined VEGF-A and VEGFR-2 concentrations in plasma: diagnostic and prognostic implications in patients with advanced NSCLC. *Lung Cancer*, 2011. **74**(2): p. 326-31.
423. Graves, E.E., A. Maity, and Q.T. Le, The tumor microenvironment in non-small-cell lung cancer. *Semin Radiat Oncol*, 2010. **20**(3): p. 156-63.
424. Hsieh, A.C. and M.M. Moasser, Targeting HER proteins in cancer therapy and the role of the non-target HER3. *Br J Cancer*, 2007. **97**(4): p. 453-7.
425. Koumakpayi, I.H., et al., Low nuclear ErbB3 predicts biochemical recurrence in patients with prostate cancer. *BJU Int*, 2007. **100**(2): p. 303-9.
426. Muller-Tidow, C., et al., Identification of metastasis-associated receptor tyrosine kinases in non-small cell lung cancer. *Cancer Res*, 2005. **65**(5): p. 1778-82.
427. Chen, H.Y., et al., A five-gene signature and clinical outcome in non-small-cell lung cancer. *N Engl J Med*, 2007. **356**(1): p. 11-20.
428. Chen, C.H., et al., Inhibition of heregulin signaling by an aptamer that preferentially binds to the oligomeric form of human epidermal growth factor receptor-3. *Proc Natl Acad Sci U S A*, 2003. **100**(16): p. 9226-31.
429. Sithanandam, G. and L.M. Anderson, The ERBB3 receptor in cancer and cancer gene therapy. *Cancer Gene Ther*, 2008. **15**(7): p. 413-48.
430. Chen, L., H. Liu, and C. Friedman, Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, 2005. **21**(2): p. 248-56.
431. Erhardt, R.A., R. Schneider, and C. Blaschke, Status of text-mining techniques applied to biomedical text. *Drug Discov Today*, 2006. **11**(7-8): p. 315-25.
432. Haury, A.C., P. Gestraud, and J.P. Vert, The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS One*, 2011. **6**(12): p. e28210.
433. Shendure, J. and H. Ji, Next-generation DNA sequencing. *Nat Biotechnol*, 2008. **26**(10): p. 1135-45.

434. Pariset, L., et al., Microarrays and high-throughput transcriptomic analysis in species with incomplete availability of genomic sequences. *N Biotechnol*, 2009. **25**(5): p. 272-9.
435. Cooper, G.M., et al., Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet*, 2008. **40**(10): p. 1199-203.
436. Abeel, T., et al., Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 2010. **26**(3): p. 392-8.

8 List of publications

1. **Zhang JX**, Han BC, Wei XN, C.Y. Tan, Y.Y. Jiang, Chen YZ. A two-step Target Binding and Selectivity Support Vector Machines Approach for Virtual Screening of Dopamine Receptor Subtype-selective Ligands. ***PLoS ONE*** 7(6): e39076. doi:10.1371/ journal.pone.0039076 (2012)

2. **Zhang JX**, J Jia, Ma XH, Han BC, Wei XN, C.Y. Tan, Y.Y. Jiang, Chen YZ. Analysis of bypass signaling in EGFR pathway and profiling of bypass genes for predicting response to anticancer EGFR tyrosine kinase inhibitors. ***Mol. BioSyst***., Advance Article, DOI: 10.1039/C2MB25165E. (2012)

3. XY Cheng; WJ Huang; SC Hu; HL Zhang; H Wang; **JX Zhang**; HH Lin; YZ Chen; Q Zou; ZL Ji. A Global Characterization and Identification of Multifunctional Enzymes. ***PLoS ONE*** 7(6): e38979. doi:10.1371/journal.pone.0038979. (2012)

4. F. Zhu, Z. Shi, C. Qin, L. Tao, X. Liu, F. Xu, L. Zhang, Y. Song, X.H. Liu, **J.X. Zhang**, B.C. Han, P. Zhang and Y.Z. Chen. Therapeutic Target Database Update 2012: A Resource for Facilitating Target-Oriented Drug Discovery. ***Nucleic Acids Res.*** 40(D1):D1128-D1136 (2012).

5. Wei XN, Han BC, **Zhang JX**, Liu XH, Tan CY, Jiang YY, Low BC, Tidor B, Chen YZ. An Integrated Mathematical Model of Thrombin-, Histamine-and VEGF-Mediated Signalling in Endothelial Permeability. ***BMC Syst Biol.*** Jul 15;5(1):112 (2011).

6. X. Liu, F. Zhu, X.H. Ma, L, Tao, **J.X. Zhang**, S.Y. Yang, Y.C. Wei and Y.Z. Chen. The Therapeutic Target Database: an internet resource for the primary targets of approved,

clinical trial and experimental drugs. ***Expert Opin Ther Targets.*** 15(8):903-12 (2011).

7: X.H. Liu, H.Y. Song, **Zhang J.X.,** B.C. Han, X.N. Wei, X.H. Ma, W.K. Chui, Y.Z. Chen. Identifying Novel Type ZBGs and Non-hydroxamate HDAC Inhibitors Through a SVM Based Virtual Screening Approach***. Mol Inf.*** 29(5): 407-20(2010)

8: Liu XX, **Zhang J.X**., Ni F, Dong X, Han BC, Han DX, Ji ZL, Zhao YF. Genome wide exploration of the origin and evolution of amino acids. ***BMC Evol Biol.*** 2010 Mar 15;10:77. (2010)

9 Appendices

Appenix A:

Table 1 Chemical databases

Company name	Link	Number of compounds	Description
4SC	www.4sc.de	5,000,000	Virtual library; small-molecule drug candidates
ACB BLOCKS	www.acbblocks.com/acb/bblocks.html	90,000	Building blocks for combinatorial chemistry
Advanced ChemTech	http://triton.peptide.com/index.php	18,000	OmniProbeTM: peptide libraries; 8000 tripeptide, 10,000 tetrapeptide
Advanced SynTech	www.advsyntech.com/omnicore.htm	170,000	Targeted libraries: protease, protein kinase, GPCR, steroid mimetics, antimicrobials
Ambinter	ourworld.compuserve.com/homepages/ambinter/Mole.htm	1,750,000	Combinatorial and parallel chemistry, building blocks, HTS
Asinex	www.asinex.com/prod/index.html	150,000	Platinum collection: drug-like compounds
Asinex		250,000	Gold collection: drug-like compounds
Asinex		5009	Targeted libraries: GPCR (16 different targets)
Asinex		4307	Kinase-targeted library (11 targets)
Asinex		1629	Ion-channel targeted (4 targets)
Asinex		2987	Protease-targeted library (5 targets)
Asinex		1,200,000	Combinatorial constructor
BioFocus	www.biofocus.com/pages/drug__discovery.mhtml	100,000	Diverse primary screening compounds
BioFocus		~16,000	SoftFocus: kinase target-directed libraries

BioFocus		~10,000	SoftFocus: GPCR target-directed libraries
CEREP	www.cerep.fr/cerep/users/pages/ProductsServices/Odyssey.asp	>16,000	Odyssey II library: diverse and unique discovery library; more than 350 chemical families
CEREP		5000	GPCR-focused library (21 targets)
Chemical Diversity	www.chemdiv.com/discovery/downloads/	>750,000	Leadlike compounds for bioscreening
ChemStar	www.chemstar.ru/page4.htm	60,260	High-quality organic compounds for screening
ChemStar		>500,000	Virtual database of organic compounds
COMBI-BLOCKS	www.combi-blocks.com	908	Combinatorial building blocks
ComGenex	www.comgenex.hu/cgi-bin/inside.php?in=products&l_id=compound	260,000	“Pharma relevant”, discrete structures for multitarget screening purposes
ComGenex		240	GPCR library
ComGenex		2000	Cytotoxic discovery library: very toxic compounds suitable for anticancer and antiviral discovery research
ComGenex		5000	Low-Tox MeDiverse: druglike, diverse, nontoxic discovery library
ComGenex		10,000	MeDiverse Natural: natural product like compounds
EMC microcolection	www.microcollections.de/catalogue_compunds.htm#	30,000	Highly diverse combinatorial compound collections for lead discovery
InterBioScreen	www.ibscreen.com/products.shtml	350,000	Synthetic compounds
InterBioScreen		40,000	Natural compounds
Maybridge plc	www.maybridge.com/html/m_company.htm	60,000	Organic druglike compounds
Maybridge plc		13,000	Building blocks
MDDR	http://www.symyx.com/products/databases/bioactivity/mddr/index.jsp	180,000	MDL Drug Data Report database

MicroSource Discovery Systems, Inc.	www.msdiscovery.com/download.html	2000	GenPlus: collection of known bioactive compounds NatProd: collection of pure natural products
Nanosyn	www.nanosyn.com/thankyou.shtml	46,715	Pharma library
Nanosyn		18,613	Explore library
Pharmacopeia Drug Discovery, Inc.	www.pharmacopeia.com/dcs/order_form.html	N/A	Targeted library: GPCR and kinase
Polyphor	www.polyphor.com	15,000	Diverse general screening library
PubChem	pubchem.ncbi.nlm.nih.gov	>16,000000	PubChem database
Sigma-Aldrich	www.sigmaaldrich.com/Area_of_Interest/Chemistry/Drug_Discovery/Assay_Dev_and_Screening/Compound_Libraries/Screening_Compounds.html	90,000	Diverse library of drug-like compounds, selected based on Lipinski Rule of Five
Specs	www.specs.net	240,000	Diverse library
Specs		10,000	World Diversity Set: pre-plateled library
Specs		6000	Building blocks
Specs		500	Natural products (diverse and unique)
TimTec	www.timtec.net	>160,000	Compound libraries and building blocks
Tranzyme Pharma	www.tranzyme.com/drug_discovery.html	25,000	HitCREATE library: macrocycles library
Tripos	www.tripos.com/sciTech/researchCollab/chemCompLib/lqCompound/index.html	80,000	LeadQuest compound libraries

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

Screening task	Compounds screened		Method and reference of reported study	Molecular descriptors	Compounds in training set (No of positives / No of negatives)	Compounds selected		Known hits selected			
	No of compounds	No of known hits included				No of compounds selected	Percentage of screened compounds selected	No of hits selected	Yield	Hit rates	Enrichment factor
COX2 inhibitors	2.5M	22	SVM [70]	Molecular fingerprints	94/200K	2,500	0.1%	18	81%	0.7%	795
	25,300	25	SVM+ BKD [71]	DRAGON descriptors	125/5035	506	2%	20	80%	3.9%	39.5
COX inhibitors	102,514	536	BKD [72, 73]	Extended connectivity fingerprints	100/400	5125	5%	76	14.3%	1.4%	2.7
	98,435	536	CKD [61]	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 [61]	Pipeline pilot	100/4000	984	1%	240	44.7%	24.4%	44.5
Thrombin inhibitors	2.5M	46	SVM [70]	Molecular fingerprints	188/200K	11,250	0.45%	25	55%	0.2%	108.7
	102,514	703	BKD [72, 73]	Extended connectivity fingerprints	100/400	5125	5%	367	52,3%	7.1%	10.3
	98,435	703	CKD [61]	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 [61]	Pipeline pilot	100/4000	984	1%	381	54.2%	38.9%	54.0
Protease inhibitors	171,726	118	SVM [74]	Extended connectivity fingerprints	228/4200	1717	1%	26	22%	1.5%	21.8
			LMNB [74]					19	16%	1%	14.5
Chemokine receptor	171,560	128	SVM [74]	Extended connectivity	258/4199	1716	1%	70	55%	4.1%	54.9

antagonists			LMNB [72, 75]	fingerprints				68	53%	3.9%	52.3
5HT3 antagonists	102,514	652	BKD [72, 73]	Extended connectivity fingerprints	100/400	5125	5%	236	36.3%	4.6%	7.2
	98,435	852	CKD [61]	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 [61]	Pipeline pilot	100/4000	984	1%	529	62.1%	54.0%	62.1
5HT1A antagonists	102,514	727	BKD [72, 73]	Extended connectivity fingerprints	100/400	5125	5%	224	30.9%	4.3%	6.1
	98,435	727	CKD [61]	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 [61]	Pipeline pilot	100/4000	984	1%	319	43.9%	32.6%	44.0
5HT reuptake inhibitors	102,514	259	BKD [72, 73]	Extended connectivity fingerprints	100/400	5125	5%	65	25%	1.2%	4.7
	98,435	259	CKD [61]	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 [61]	Pipeline pilot	100/4000	984	1%	137	52.9%	14.0%	53.8
D2 antagonists	102,514	295	BKD [72, 73]	Extended connectivity fingerprints	100/400	5125	5%	90	30.6%	1.7%	5.9
	98,435	295	CKD [61]	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 [61]	Pipeline pilot	100/4000	984	1%	137	46.4%	14.0%	53.8
Rennin inhibitors	102,514	1030	BKD [72, 73]	Extended connectivity fingerprints	100/400	5125	5%	972	94.4%	18.9%	18.9
	98,435	1030	CKD [61]	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 [61]	Pipeline pilot	100/4000	984	1%	710	68.9%	72.4%	69.0
Angiotesin II AT1 antagonists	102,514	843	BKD [72, 73]	Extended connectivity fingerprints	100/400	5125	5%	776	92.1%	15.1%	18.4
	98,435	843	CKD [61]	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 [61]	Pipeline pilot	100/4000	984	1%	384	45.6%	39.2%	45.6
Substance P antagonists	102,514	1146	BKD [72, 73]	Extended connectivity fingerprints	100/400	5125	5%	378	33%	7.3%	6.5
	98,435	1146	CKD [61]	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 [61]	Pipeline pilot	100/4000	984	1%	509	44.4%	51.9%	44.4
HIV protease inhibitors	102,514	650	BKD [72, 73]	Extended connectivity fingerprints	100/400	5125	5%	377	58%	7.3%	11.5
	98,435	650	CKD [61]	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 [61]	Pipeline pilot	100/4000	984	1%	355	54.6%	36.2%	54.9
Protein kinase C inhibitors	102,514	353	BKD [72, 73]	Extended connectivity fingerprints	100/400	5125	5%	81	23.1%	1.5%	4.4
	98,435	353	CKD [61]	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 [61]	Pipeline pilot	100/4000	984	1%	206	58.3%	21.0%	58.3
MAO inhibitors	101,437	1166	BKD [76]	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 [61]	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 [61]	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 [61]	Pipeline pilot	100/4000	984	1%	192	69.3%	19.5%	69.7
				ECFP4	100/4000	984	1%	244	88.2%	27.3%	97.6
Aldose reductase inhibitors	98,435	782	CKD [61]	Pipeline pilot	100/4000	984	1%	436	55.8%	44.3%	56.1
				ECFP4	100/4000	984	1%	665	85.0%	67.6%	85.5
Reverse transcriptase inhibitors	98,435	419	CKD [61]	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 [61]	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 [61]	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 [71]	DRAGON descriptors	125/5035	506	2%	18	72%	3.5%	35.4
FXa inhibitors	25,300	25	SVM+ BKD [71]	DRAGON descriptors	125/5035	506	2%	21	84%	4.1%	N/A
PDE5 inhibitors	50,000	19	RO5+ DS [77]	Pharmacophore and macroscopic descriptors	130/10K	1821	3.6%	11	57.8%	0.6%	15.8
	25,300	25	SVM+ BKD [71]	DRAGON descriptors	125/5035	506	2%	21	84%	4.1%	41.5
Alpha1A AR antagonists	25,300	25	SVM+ BKD [71]	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 3 Performance of docking methods in virtual screening test for identifying inhibitors, agonists and substrates of proteins of pharmaceutical relevance; the relevant literature references are given in the method column**.**

Screening task	**Compounds screened**		**Method and reference of reported study**	**No of pre-docking selected compounds**	**Docking cut-off**	**Compounds selected**		**Known hits selected**			
	No of compounds	**No of known hits included**				**No of compounds selected**	**Percentage of screened compounds selected**	**No of hits selected**	**Yield**	**Hit rates**	**Enrichment factor**
Factor Xa inhibitors	2M	630	AUTODOCK + pre-docking RO5 and EA screen [78]	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 [79]	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 [80]	<400K	N/A	35	0.0087%	4	N/A	11.4%	N/A
Thyroid hormone receptor antagonists	250K	>14	ICM VLS module (Molsoft) [81] + 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 [82]	165,581	Top-500 + Top-500	889	0.38%	127	N/A	14.3%	N/A
	141K	10	GOLD + elements and chemical group	<141K	Top-2%	<2820	<2.5%	8	80%	<0.28%	39.4

			screen [83]								
BCL-2 inhibitors	206,876	>1	DOCK3.5 + non-peptidic screen [84]	<206,876	Top-500	35	0.017%	1	N/A	2.9%	N/A
HIV-1 protease inhibitors	141K	5	GLIDE + elements and chemical group screen [83]	<141K	Top-5%	<7050	<5%	1	20%	<0.014%	4.6
HDM2 inhibitors	141K	14	DOCK + elements and chemical group screen [83]	<141K	Top-5%	<7050	<5%	4	28.6%	<0.056%	5.7
UPA inhibitors	141K	10	GOLD + elements and chemical group screen [83]	<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 [85]	22,950	Top-300	300	0.21%	38	N/A	N/A	N/A
Thrombin inhibitors	141K	10	GLIDE + elements and chemical group screen [83]	<141K	Top-2%	<2820	<2.5%	3	30%	<0.11%	15.5
	133.8K	760	FlexX + Similarity [86]	<133.8K	Top-1%	1338	1%	231	29.3%	17.3%	30.5
DHFR inhibitors	135K	165	DOCK3.5.54 applied to holo form [87]	135K	Top-1% of 50k docked	1350	1%	47	25%	3.4%	27.8
			DOCK3.5.54 applied to appo form [87]	135K	Top-1% of 100k docked	1000	1%	16	9.7%	1.6%	13.1
Neutral endopeptidase	135K	356	DOCK3.5.54 [87]	135K	Top-1% of	1255	0.74%	3	0.8%	0.24%	~1

inhibitors					125.5K docked						
Thrombin inhibitors	135K	788	DOCK3.5.54 [87]	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 [87]	135K	Top-1% of 54K docked	540	0.4%	49	26.5%	9.1%	66.4
Phospholipase C inhibitors	135K	25	DOCK3.5.54 [87]	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 [87]	135K	Top-5% of database	4500	3.3%	10	2.8%	0.22%	~1
			DOCK3.5.54 applied to appo form [87]	135K	Top-5% of database	4500	3.3%	5	1.4%	0.11%	<1
	133.8K	59	FlexX + Similarity [86]	<133.8K	Top-1%	1338	1%	13	22%	0.97%	22.0
Acetylcholinesterase inhibitors	135K	637	DOCK3.5.54 applied to holo form [87]	135K	Top-1% of 77K docked	770	0.57%	49	7.7%	6.4%	13.6
			DOCK3.5.54 applied to appo form [87]	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 [86]	<133.8K	Top-1%	1338	1%	35	3.4%	2.6%	3.4

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

Screening task	**Compounds screened**		**Method and reference of reported study**	**Compounds selected**		**Known hits selected**			
	No of compounds	**No of known hits included**		**No of compounds selected**	**Percentage of screened compounds selected**	**No of hits selected**	**Yield**	**Hit rates**	**Enrichment factor**
ACE inhibitors	3.8M	55	Pharmacophore [88]	1M	26%	39	70.1%	0.0039%	2.8
	3.8M	55	Structure-based pharmacophore [89]	91K	2.4%	6	10.9%	0.0066%	4.6
11β-hydroxysteroid dehydrogenase 1 inhibitors	1.77M	144	Pharmacophore [55]	20.3K	1.15%	17	11.8%	0.084%	10.3
Rhinovirus 3C protease inhibitors	380K	30	Pharmacophore [56]	6,917	1.82%	23	76.7%	0.33%	41.8

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

Screening task	Compounds screened		Method and reference of reported study	Compounds selected		Known hits selected			
	No of compounds	No of known hits included		No of compounds selected	Percentage of screened compounds selected	No of hits selected	Yield	Hit rates	Enrichment factor
ACE inhibitors	344.5K	490	Hierachical k-means [54]	5590	1.6%	246	50.2%	4.4%	31.2
			NIPALSTREE [54]	8174	2.4%	188	38.4%	2.3%	16.2
			Hierachical k-means + NIPALSTREE disjunction [54]	12240	3.6%	306	62.4%	2.5%	17.6
			Hierachical k-means + NIPALSTREE conjunction [54]	1662	0.48%	128	26.1%	7.7%	54
COX inhibitors	344.5K	1556	Hierachical k-means [54]	15322	4.4%	761	48.9%	5.0%	11
			NIPALSTREE [54]	22321	6.5%	625	40.2%	2.8%	6.16
			Hierachical k-means + NIPALSTREE disjunction [54]	33793	9.8%	980	63.0%	2.9%	6.42
			Hierachical k-means + NIPALSTREE conjunction [54]	3980	1.2%	406	26.1%	10.2%	22.6
Adrenoceptor ligand	344.5K	542	Hierachical k-means [54]	21285	6.2%	298	55.0%	1.4%	8.99
			NIPALSTREE [54]	28125	8.2%	270	49.8%	0.96%	6.14
			Hierachical k-means + NIPALSTREE disjunction [54]	42365	12.3%	394	72.7%	0.93%	5.93
			Hierachical k-means + NIPALSTREE conjunction [54]	6692	1.9%	174	32.1%	2.6%	16..3
Glucocorticoid receptor ligand	344.5K	91	Hierachical k-means [54]	3750	1.1%	27	29.7%	0.72%	27..3
			NIPALSTREE [54]	3469	1.0%	17	18.7%	0.49%	18.7
			Hierachical k-means + NIPALSTREE disjunction [54]	7317	2.1%	30	33.0%	0.41%	15.6

			Hierachical k-means + NIPALSTREE conjunction [54]	538	0.16%	14	15.4%	2.6%	98
GABA receptor ligand	344.5K	478	Hierachical k-means [54]	10000	2.9%	110	23%	1.1%	7.97
			NIPALSTREE [54]	17143	5.0%	84	17.6%	0.49%	3.51
			Hierachical k-means + NIPALSTREE disjunction [54]	24265	7.0%	165	34.5%	0.68%	4.86
			Hierachical k-means + NIPALSTREE conjunction [54]	2636	0.77%	29	6.1%	1.1%	7.77

Appendix B:

Table 1 The 148 gefitinib response biomarkers selected by our SVM-RFE method from the 38 and 6 gefitinib resistant and sensitive NSCLC cell-lines, the biomarkers selected by a previously published study or as the gefitinib target or bypass gene are marked in the Table.

SVM-RFE selected biomarker		**Biomarker commonly selected by a previous published study**		**Biomarker as gefitinib target or bypass gene**	
Probeset ID	Gene Symbol	Gene Description	Selected by the study of [371]	Selected by the study of [293]	Bypass gene
212895_s_at	ABR	active BCR-related gene			
202982_s_at	ACOT1	acyl-CoA thioesterase 2			
218795_at	ACP6	acid phosphatase 6, lysophosphatidic			
202666_s_at	ACTL6A	actin-like 6A			
219199_at	AFF4	AF4/FMR2 family, member 4			
202054_s_at	ALDH3A2	aldehyde dehydrogenase 3 family, member A2			
221825_at	ANGEL2	angel homolog 2 (Drosophila)			
206200_s_at	ANXA11	annexin A11			
221653_x_at	APOL2	apolipoprotein L, 2			
203025_at	ARD1A	ARD1 homolog A, N-acetyltransferase (S. cerevisiae)			
219335_at	ARMCX5	armadillo repeat containing, X-linked 5			
207076_s_at	ASS1	argininosuccinate synthetase 1			
209492_x_at	ATP5I	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit E			
218631_at	AVPI1	arginine vasopressin-induced 1			
203304_at	BAMBI	BMP and activin membrane-bound inhibitor homolog (Xenopus laevis)			
202331_at	BCKDHA	branched chain keto acid dehydrogenase E1, alpha polypeptide			
201101_s_at	BCLAF1	BCL2-associated transcription factor 1			
211715_s_at	BDH1	3-hydroxybutyrate dehydrogenase, type 1			
218792_s_at	BSPRY	B-box and SPRY domain containing		√	
218462_at	BXDC5	brix domain containing 5			
219240_s_at	C10orf88	chromosome 10 open reading frame 88			
217969_at	C11orf2	chromosome 11 open reading frame2			
219099_at	C12orf5	chromosome 12 open reading frame 5			
218940_at	C14orf138	chromosome 14 open reading frame 138			
218183_at	C16orf5	chromosome 16 open reading frame 5			

219260_s_at	C17orf81	chromosome 17 open reading frame 81			
212574_x_at	C19orf6	chromosome 19 open reading frame 6			
213528_at	C1orf156	chromosome 1 open reading frame 156			
220477_s_at	C20orf30	chromosome 20 open reading frame 30			
219329_s_at	C2orf28	chromosome 2 open reading frame 28			
219008_at	C2orf43	chromosome 2 open reading frame 43			
213148_at	C2orf72	chromosome 2 open reading frame 72			
218646_at	C4orf27	chromosome 4 open reading frame 27			
206016_at	CCDC22	coiled-coil domain containing 22			
218026_at	CCDC56	coiled-coil domain containing 56			
213743_at	CCNT2	cyclin T2			
204306_s_at	CD151	CD151 molecule (Raph blood group)			
204693_at	CDC42EP1	CDC42 effector protein (Rho GTPase binding) 1			
203493_s_at	CEP57	centrosomal protein 57kDa			
212228_s_at	COQ9	coenzyme Q9 homolog (S. cerevisiae)			
206918_s_at	CPNE1	copine I			
201983_s_at	EGFR	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)			√
201313_at	ENO2	enolase 2 (gamma, neuronal)			
217941_s_at	ERBB2IP	erbb2 interacting protein			
202454_s_at	ERBB3	v-erb-b2 erythroblastic leukemia viral oncogene homolog 3 (avian)		√	√
218481_at	EXOSC5	exosome component 5			
207813_s_at	FDXR	ferredoxin reductase			
219901_at	FGD6	FYVE, RhoGEF and PH domain containing 6			
207822_at	FGFR1	fibroblast growth factor receptor 1			√
206095_s_at	FUSIP1	FUS interacting protein (serine/arginine-rich) 1			
203987_at	FZD6	frizzled homolog 6 (Drosophila)			
218313_s_at	GALNT7	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 7 (GalNAc-T7)			
206920_s_at	GLE1	GLE1 RNA export mediator homolog (yeast)			
201501_s_at	GRSF1	G-rich RNA sequence binding factor 1			
201470_at	GSTO1	glutathione S-transferase omega 1			
201007_at	HADHB	hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzy			

		me A thiolase/enoyl-Coenzyme A hydratase (trifunctional protein), beta subunit			
218460_at	HEATR2	HEAT repeat containing 2			
210982_s_at	HLA-DRA	major histocompatibility complex, class II, DR alpha			
202854_at	HPRT1	hypoxanthine phosphoribosyltransferase 1			
212221_x_at	IDS	iduronate 2-sulfatase			
221548_s_at	ILKAP	integrin-linked kinase-associated serine/threonine phosphatase 2C			
213392_at	IQCK	IQ motif containing K			
217938_s_at	KCMF1	potassium channel modulatory factor 1			
212303_x_at	KHSRP	KH-type splicing regulatory protein			
201776_s_at	KIAA0494	KIAA0494			
217906_at	KLHDC2	kelch domain containing 2			
206316_s_at	KNTC1	kinetochore associated 1			
202594_at	LEPROTL1	leptin receptor overlapping transcript-like 1			
209205_s_at	LMO4	LIM domain only 4			
201569_s_at	LOC100131861	sorting and assembly machinery component 50 homolog (S. cerevisiae)			
220130_x_at	LTB4R2	leukotriene B4 receptor 2			
203497_at	MED1	mediator complex subunit 1			
211599_x_at	MET	met proto-oncogene (hepatocyte growth factor receptor)			√
209124_at	MYD88	myeloid differentiation primary response gene (88)			
219946_x_at	MYH14	myosin, heavy chain 14		√	
37005_at	NBL1	neuroblastoma, suppression of tumorigenicity 1			
200854_at	NCOR1	nuclear receptor co-repressor 1			
203245_s_at	NCRNA00094	non-protein coding RNA 94			
219726_at	NLGN3	neuroligin 3			
205895_s_at	NOLC1	nucleolar and coiled-body phosphoprotein 1			
214661_s_at	NOP14	NOP14 nucleolar protein homolog (yeast)			
209628_at	NXT2	nuclear transport factor 2-like export factor 2			
201282_at	OGDH	oxoglutarate (alpha-ketoglutarate) dehydrogenase (lipoamide)			
212858_at	PAQR4	progestin and adipoQ receptor family member IV			
211048_s_at	PDIA4	protein disulfide isomerase family A, member 4			
202464_s_at	PFKFB3	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3			
219235_s_at	PHACTR	phosphatase and actin regulator 4			

	4				
217954_s_at	PHF3	PHD finger protein 3			
211668_s_at	PLAU	plasminogen activator, urokinase			
206080_at	PLCH2	phospholipase C, eta 2			
203735_x_at	PPFIBP1	PTPRF interacting protein, binding protein 1 (liprin beta 1)			
201300_s_at	PRNP	prion protein			
201705_at	PSMD7	proteasome (prosome, macropain) 26S subunit, non-ATPase, 7			
219938_s_at	PSTPIP2	proline-serine-threonine phosphatase interacting protein 2			
221808_at	RAB9A	RAB9A, member RAS oncogene family			
205037_at	RABL4	RAB, member of RAS oncogene family-like 4			
201039_s_at	RAD23A	RAD23 homolog A (S. cerevisiae)			
210621_s_at	RASA1	RAS p21 protein activator (GTPase activating protein) 1			
212646_at	RFTN1	raftlin, lipid raft linker 1			
218564_at	RFWD3	ring finger and WD repeat domain 3			
218323_at	RHOT1	ras homolog gene family, member T1			
214700_x_at	RIF1	RAP1 interacting factor homolog (yeast)			
201823_s_at	RNF14	ring finger protein 14			
208540_x_at	S100A11	S100 calcium binding protein A11 pseudogene			
201747_s_at	SAFB	scaffold attachment factor B			
203455_s_at	SAT1	spermidine/spermine N1-acetyltransferase 1			
201339_s_at	SCP2	sterol carrier protein 2			
202657_s_at	SERTAD2	SERTA domain containing 2			
216457_s_at	SF3A1	splicing factor 3a, subunit 1, 120kDa			
200753_x_at	SFRS2	splicing factor, arginine/serine-rich 2			
218878_s_at	SIRT1	sirtuin (silent mating type information regulation 2 homolog) 1 (S. cerevisiae)			
205896_at	SLC22A4	solute carrier family 22 (organic cation/ergothioneine transporter), member 4			
221020_s_at	SLC25A32	solute carrier family 25, member 32			
218041_x_at	SLC38A2	solute carrier family 38, member 2			
203579_s_at	SLC7A6	solute carrier family 7 (cationic amino acid transporter, y+ system), member 6			
202043_s_at	SMS	spermine synthase			
207390_s_at	SMTN	smoothelin			
205443_at	SNAPC1	small nuclear RNA activating complex, polypeptide 1, 43kDa			

201221_s_at	SNRNP70	small nuclear ribonucleoprotein 70kDa (U1)			
201522_x_at	SNRPN	small nuclear ribonucleoprotein polypeptide N			
203217_s_at	ST3GAL5	ST3 beta-galactoside alpha-2,3-sialyltransferase 5			
205339_at	STIL	SCL/TAL1 interrupting locus			
202786_at	STK39	serine threonine kinase 39 (STE20/SPS1 homolog, yeast)		√	
202260_s_at	STXBP1	syntaxin binding protein 1			
205759_s_at	SULT2B1	sulfotransferase family, cytosolic, 2B, member 1			
202384_s_at	TCOF1	Treacher Collins-Franceschetti syndrome 1			
220407_s_at	TGFB2	transforming growth factor, beta 2			
219580_s_at	TMC5	transmembrane channel-like 5		√	
219005_at	TMEM59L	transmembrane protein 59-like			
220431_at	TMPRSS11E	transmembrane protease, serine 11E			
206907_at	TNFSF9	tumor necrosis factor (ligand) superfamily, member 9			
207196_s_at	TNIP1	TNFAIP3 interacting protein 1			
202734_at	TRIP10	thyroid hormone receptor interactor 10			
209109_s_at	TSPAN6	tetraspanin 6			
213058_at	TTC28	tetratricopeptide repeat domain 28			
211285_s_at	UBE3A	ubiquitin protein ligase E3A			
219960_s_at	UCHL5	ubiquitin carboxyl-terminal hydrolase L5			
201903_at	UQCRC1	ubiquinol-cytochrome c reductase core protein I			
201831_s_at	USO1	USO1 homolog, vesicle docking protein (yeast)			
202664_at	WIPF1	WAS/WASL interacting protein family, member 1			
201760_s_at	WSB2	WD repeat and SOCS box-containing 2			
204022_at	WWP2	WW domain containing E3 ubiquitin protein ligase 2			
221423_s_at	YIPF5	Yip1 domain family, member 5			
212787_at	YLPM1	YLP motif containing 1			
201531_at	ZFP36	zinc finger protein 36, C3H type, homolog (mouse)			
203730_s_at	ZKSCAN5	zinc finger with KRAB and SCAN domains 5			
203247_s_at	ZNF24	zinc finger protein 24			
206829_x_at	ZNF430	zinc finger protein 430			

Table 2 The 65 Erlotinib response biomarkers selected by our SVM-RFE method from the 46 and 7 Erlotinib resistant and sensitive NSCLC cell-lines, the biomarkers selected by a previously published study or as the Erlotinib target or bypass gene are marked in the Table.

SVM-RFE selected biomarker			**Biomarker commonly selected by a previous published study**	**Biomarker as Erlotinib target or bypass gene**
Probeset ID	Gene Symbol	Gene Description	Selected by the study of [328]	Bypass gene
219199_at	AFF4	AF4/FMR2 family, member 4		
202054_s_at	ALDH3A2	aldehyde dehydrogenase 3 family, member A2		
221825_at	ANGEL2	angel homolog 2 (Drosophila)		
204416_x_at	APOC1	apolipoprotein C-I		
203311_s_at	ARF6	ADP-ribosylation factor 6		
207076_s_at	ASS1	argininosuccinate synthetase 1		
214068_at	BEAN	brain expressed, associated with Nedd4		
217969_at	C11orf2	chromosome 11 open reading frame2		
221208_s_at	C11orf61	chromosome 11 open reading frame 61		
219260_s_at	C17orf81	chromosome 17 open reading frame 81		
220477_s_at	C20orf30	chromosome 20 open reading frame 30		
219329_s_at	C2orf28	chromosome 2 open reading frame 28		
213148_at	C2orf72	chromosome 2 open reading frame 72		
213322_at	C6orf130	chromosome 6 open reading frame 130		
206016_at	CCDC22	coiled-coil domain containing 22		
204306_s_at	CD151	CD151 molecule (Raph blood group)		
212648_at	DHX29	DEAH (Asp-Glu-Ala-His) box polypeptide 29		
200606_at	DSP	desmoplakin		
201983_s_at	EGFR	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)	√	√
217941_s_at	ERBB2IP	erbb2 interacting protein		
202454_s_at	ERBB3	v-erb-b2 erythroblastic leukemia viral oncogene homolog 3 (avian)		√
218481_at	EXOSC5	exosome component 5		
218898_at	FAM57A	family with sequence similarity 57, member A		
213646_x_at	FGFR3	fibroblast growth factor receptor 3		√
207966_s_at	GLG1	golgi apparatus protein 1		
203384_s_at	GOLGA1	golgi autoantigen, golgin subfamily a, 1		
206204_at	GRB14	growth factor receptor-bound protein 14	√	
201631_s_at	IER3	immediate early response 3		
221548_s_at	ILKAP	integrin-linked kinase-associated serine/threonine phosphatase 2C		
202419_at	KDSR	3-ketodihydrosphingosine reductase		
209205_s_at	LMO4	LIM domain only 4		
216908_x_at	LOC7300	RRN3 RNA polymerase I transcription		

	92	factor homolog (S. cerevisiae) pseudogene		
217543_s_at	MBTPS1	membrane-bound transcription factor peptidase, site 1		
211599_x_at	MET	met proto-oncogene (hepatocyte growth factor receptor)		√
209124_at	MYD88	myeloid differentiation primary response gene (88)		
37005_at	NBL1	neuroblastoma, suppression of tumorigenicity 1		
203245_s_at	NCRNA00094	non-protein coding RNA 94		
218036_x_at	NMD3	NMD3 homolog (S. cerevisiae)		
209628_at	NXT2	nuclear transport factor 2-like export factor 2		
221538_s_at	PLXNA1	plexin A1		
37028_at	PPP1R15A	protein phosphatase 1, regulatory (inhibitor) subunit 15A		
221808_at	RAB9A	RAB9A, member RAS oncogene family		
205037_at	RABL4	RAB, member of RAS oncogene family-like 4		
210621_s_at	RASA1	RAS p21 protein activator (GTPase activating protein) 1	√	
218323_at	RHOT1	ras homolog gene family, member T1		
208540_x_at	S100A11P	S100 calcium binding protein A11 pseudogene		
201339_s_at	SCP2	sterol carrier protein 2		
216457_s_at	SF3A1	splicing factor 3a, subunit 1, 120kDa		
200753_x_at	SFRS2	splicing factor, arginine/serine-rich 2		
205896_at	SLC22A4	solute carrier family 22 (organic cation/ergothioneine transporter), member 4		
221020_s_at	SLC25A32	solute carrier family 25, member 32		
203579_s_at	SLC7A6	solute carrier family 7 (cationic amino acid transporter, y+ system), member 6		
202043_s_at	SMS	spermine synthase		
218327_s_at	SNAP29	synaptosomal-associated protein, 29kDa		
200783_s_at	STMN1	stathmin 1/oncoprotein 18		
202260_s_at	STXBP1	syntaxin binding protein 1		
203449_s_at	TERF1	telomeric repeat binding factor (NIMA-interacting) 1		
219580_s_at	TMC5	transmembrane channel-like 5		
206907_at	TNFSF9	tumor necrosis factor (ligand) superfamily, member 9		
201546_at	TRIP12	thyroid hormone receptor interactor 12		
211758_x_at	TXNDC9	thioredoxin domain containing 9		
201649_at	UBE2L6	ubiquitin-conjugating enzyme E2L 6		
202664_at	WIPF1	WAS/WASL interacting protein family, member 1		
221423_s_at	YIPF5	Yip1 domain family, member 5		
212787_at	YLPM1	YLP motif containing 1		

Table 3 List of the 98 Lapatinib response biomarkers selected by our SVM-RFE method from the 40 and 8 Lapatinib resistant and sensitive NSCLC cell-lines, the biomarkers as the Lapatinib target or bypass gene are marked in the Table.

SVM-RFE selected biomarker			**Biomarker as Lapatinib target or bypass gene**
Probeset ID	Gene Symbol	Gene Description	
212895_s_at	ABR	active BCR-related gene	
205512_s_at	AIFM1	apoptosis-inducing factor, mitochondrion-associated, 1	
204416_x_at	APOC1	apolipoprotein C-I	
221653_x_at	APOL2	apolipoprotein L, 2	
220658_s_at	ARNTL2	aryl hydrocarbon receptor nuclear translocator-like 2	
207076_s_at	ASS1	argininosuccinate synthetase 1	
209406_at	BAG2	BCL2-associated athanogene 2	
222000_at	C1orf174	chromosome 1 open reading frame 174	
218646_at	C4orf27	chromosome 4 open reading frame 27	
218026_at	CCDC56	coiled-coil domain containing 56	
219036_at	CEP70	centrosomal protein 70kDa	
213735_s_at	COX5B	cytochrome c oxidase subunit Vb	
203368_at	CRELD1	cysteine-rich with EGF-like domains 1	
201201_at	CSTB	cystatin B (stefin B)	
205399_at	DCLK1	doublecortin-like kinase 1	
209916_at	DHTKD1	dehydrogenase E1 and transketolase domain containing 1	
209190_s_at	DIAPH1	diaphanous homolog 1 (Drosophila)	
218976_at	DNAJC12	DnaJ (Hsp40) homolog, subfamily C, member 12	
222221_x_at	EHD1	EH-domain containing 1	
201313_at	ENO2	enolase 2 (gamma, neuronal)	
210930_s_at	ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	√
217941_s_at	ERBB2IP	erbb2 interacting protein	
202454_s_at	ERBB3	v-erb-b2 erythroblastic leukemia viral oncogene homolog 3 (avian)	√
202766_s_at	FBN1	fibrillin 1	
213646_x_at	FGFR3	fibroblast growth factor receptor 3	√
214170_x_at	FH	fumarate hydratase	
215075_s_at	GRB2	growth factor receptor-bound protein 2	
201209_at	HDAC1	histone deacetylase 1	
208306_x_at	HLA-DRB4	major histocompatibility complex, class II, DR beta 4	
209417_s_at	IFI35	interferon-induced protein 35	
219209_at	IFIH1	interferon induced with helicase C domain 1	

202859_x_at	IL8	interleukin 8	
205051_s_at	KIT	v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog	
210111_s_at	KLHDC10	kelch domain containing 10	
209008_x_at	KRT8	keratin 8	
217892_s_at	LIMA1	LIM domain and actin binding 1	
209737_at	MAGI2	membrane associated guanylate kinase, WW and PDZ domain containing 2	
205192_at	MAP3K14	mitogen-activated protein kinase kinase kinase 14	
213927_at	MAP3K9	mitogen-activated protein kinase kinase kinase 9	
218440_at	MCCC1	methylcrotonoyl-Coenzyme A carboxylase 1 (alpha)	
200617_at	MLEC	malectin	
201710_at	MYBL2	v-myb myeloblastosis viral oncogene homolog (avian)-like 2	
209498_at	MYCN	v-myc myelocytomatosis viral related oncogene, neuroblastoma derived (avian)	
209124_at	MYD88	myeloid differentiation primary response gene (88)	
208754_s_at	NAP1L1	nucleosome assembly protein 1-like 1	
217286_s_at	NDRG3	NDRG family member 3	
202647_s_at	NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog	
212316_at	NUP210	nucleoporin 210kDa	
215952_s_at	OAZ1	ornithine decarboxylase antizyme 1	
209043_at	PAPSS1	3'-phosphoadenosine 5'-phosphosulfate synthase 1	
203131_at	PDGFR	platelet-derived growth factor receptor, alpha polypeptide	√
219165_at	PDLIM2	PDZ and LIM domain 2 (mystique)	
221538_s_at	PLXNA1	plexin A1	
209317_at	POLR1C	polymerase (RNA) I polypeptide C, 30kDa	
209482_at	POP7	processing of precursor 7, ribonuclease P/MRP subunit (S. cerevisiae)	
204748_at	PTGS2	prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	√
212032_s_at	PTOV1	prostate tumor overexpressed 1	
203329_at	PTPRM	protein tyrosine phosphatase, receptor type, M	
206157_at	PTX3	pentraxin-related gene, rapidly induced by IL-1 beta	
211823_s_at	PXN	paxillin	
213878_at	PYROXD1	pyridine nucleotide-disulphide oxidoreductase domain 1	
219681_s_at	RAB11FIP1	RAB11 family interacting protein 1 (class I)	
208732_at	RAB2A	RAB2A, member RAS oncogene family	
210621_s_at	RASA1	RAS p21 protein activator (GTPase	

		activating protein) 1	
208242_at	RAX	retina and anterior neural fold homeobox	
213718_at	RBM4	RNA binding motif protein 4	
205740_s_at	RBM42	RNA binding motif protein 42	
218323_at	RHOT1	ras homolog gene family, member T1	
216913_s_at	RRP12	ribosomal RNA processing 12 homolog (S. cerevisiae)	
218307_at	RSAD1	radical S-adenosyl methionine domain containing 1	
208540_x_at	S100A11P	S100 calcium binding protein A11 pseudogene	
203408_s_at	SATB1	SATB homeobox 1	
203889_at	SCG5	secretogranin V (7B2 protein)	
201339_s_at	SCP2	sterol carrier protein 2	
214016_s_at	SFPQ	splicing factor proline/glutamine-rich (polypyrimidine tract binding protein associated)	
202433_at	SLC35B1	solute carrier family 35, member B1	
204368_at	SLCO2A1	solute carrier organic anion transporter family, member 2A1	
218327_s_at	SNAP29	synaptosomal-associated protein, 29kDa	
205443_at	SNAPC1	small nuclear RNA activating complex, polypeptide 1, 43kDa	
204729_s_at	STX1A	syntaxin 1A (brain)	
210580_x_at	SULT1A3	sulfotransferase family, cytosolic, 1A, phenol-preferring, member 3	
203167_at	TIMP2	TIMP metallopeptidase inhibitor 2	
212165_at	TMEM183A	transmembrane protein 183A	
202688_at	TNFSF10	tumor necrosis factor (ligand) superfamily, member 10	
206907_at	TNFSF9	tumor necrosis factor (ligand) superfamily, member 9	
214550_s_at	TNPO3	transportin 3	
203050_at	TP53BP1	tumor protein p53 binding protein 1	
202734_at	TRIP10	thyroid hormone receptor interactor 10	
215111_s_at	TSC22D1	TSC22 domain family, member 1	
217979_at	TSPAN13	tetraspanin 13	
46167_at	TTC4	tetratricopeptide repeat domain 4	
211285_s_at	UBE3A	ubiquitin protein ligase E3A	
202316_x_at	UBE4B	ubiquitination factor E4B (UFD2 homolog, yeast)	
220419_s_at	USP25	ubiquitin specific peptidase 25	
205139_s_at	UST	uronyl-2-sulfotransferase	
201531_at	ZFP36	zinc finger protein 36, C3H type, homolog (mouse)	
218645_at	ZNF277	zinc finger protein 277	
218735_s_at	ZNF544	zinc finger protein 544	